Ahmet Kondoz · Tasos Dagiuklas *Editors*

3D Future Internet Media

Springer

3D Future Internet Media

Ahmet Kondoz • Tasos Dagiuklas
Editors

3D Future Internet Media

Springer

Editors
Ahmet Kondoz
Faculty Engineering and Physical Sciences
Centre for Vision Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, United Kingdom

Tasos Dagiuklas
Department of Computer Science
Hellenic Open University
Patras, Greece

ISBN 978-1-4614-8372-4 ISBN 978-1-4614-8373-1 (eBook)
DOI 10.1007/978-1-4614-8373-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013950183

© Springer Science+Business Media New York 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Hemantha Kodikara Arachchi I-Lab Multimedia Communication Research, University of Surrey, Guildford, UK

Pedro Assunção Instituto de Telecomunicações/Instituto Politécnico de Leiria, Delegação de Leiria, Portugal

Marcus Barkowsky LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, Nantes, France

Riccardo Bassoli Instituto de Telecomunicações, Campus Universitário Santiago, Aveiro, Portugal

Konstantinos Birkos University of Patras, Patras, Greece

Patrick Le Callet LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, Nantes, France

Tasos Dagiuklas Department of Computer Science, Hellenic Open University, Patras, Greece

Carl James Debono Department of Communications and Computer Engineering, University of Malta, Msida, Malta

Athanasios M. Demiris Micro2gen Ltd., Technology Park NCSR Demokritos, Agia Paraskevi, Greece

Gabriel Dosso VITEC Multimedia, Châtillon, France

Erhan Ekmekcioglu I-Lab Multimedia Communications Research Group, University of Surrey, Guildford, Surrey, UK

Ignacio Digón Escudero Telefónica I+D, Madrid, Spain

Sérgio Faria Instituto de Telecomunicações/Instituto Politécnico de Leiria, Delegação de Leiria, Portugal

Georgios Gardikis NCSR "Demokritos", Institute of Informatics and Telecommunications, Aghia Paraskevi, Attiki, Greece

Michael Grafl Alpen-Adria-Universitaet Klagenfurt, Universitätsstraße, Klagenfurt, Austria

Chaminda T.E.R. Hewage Kingston University, London, UK

Chungeun Kim I-Lab Multimedia and DSP Research Group, Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

Ahmet Kondoz Faculty Engineering & Physical Sciences, Centre for Vision Speech and Signal, University of Surrey, Guildford, UK

Athanasios Kordelas Department of Electrical & Computer Engineering, University of Patras, Patras, Greece

Jing Li LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, Nantes, France

Evariste Logota Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Asimakis Lykourgiotis Department of Electrical and Computer Engineering, University of Patras, Rio, Hellas

Hugo Marques Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal

Maria G. Martini Kingston University, London, UK

Brian Walter Micallef Department of Communications and Computer Engineering, University of Malta, Msida, Malta

Moustafa M. Nasralla Kingston University, London, UK

Evangelos Pallis TEI of Crete, Department of Applied Informatics and Multimedia, Heraklion, Crete, Greece

Fernando Pascual Blanco Telefónica I+D, Madrid, Spain

Luís Pinto Instituto de Telecomunicações/Instituto Politécnico de Leiria, Delegação de Leiria, Portugal

Ilias Politis Department of Electrical & Computer Engineering, University of Patras, Patras, Greece

Jonathan Rodriguez Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Manuel Nuñez Sanz Telefónica I+D, Madrid, Spain

Nicolas Tizon VITEC Multimedia, Châtillon, France

Chapter 1
Introduction

Ahmet Kondoz and Tasos Dagiuklas

There have been many significant advances in 3D media technologies in terms of capturing, representing, coding, transmitting, and visualizing for 3D displays. 3D media have been evolving in various areas, covering diverse market segments such as professional (e.g., scientific, medical, education, and training) and entertainment (3D interactive gaming, broadcasting, social networking, etc.) sectors. These new technologies provide the ability to design and develop applications ranging from virtual collaborative environments (e.g., multimodal interactions, seamless application on generation) to edutainment (e.g., 3D telepresence, combination of the educational content along with entertainment).

Since the Internet has grown beyond its original design objectives due to the increasing demand for performance, availability, scalability, security, and reliability, it progressively reaches a set of fundamental technological (evolution in wireless and mobile networking technologies) and operational limitations (e.g., exhausting the number of possible IP addresses). The Internet was designed for purposes that bear little resemblance to today's usage scenarios and related traffic patterns.

Several organizations have been working towards the development of the Future Internet. Many research projects have been established worldwide. For example, in Europe the FIRE [1], in the USA the FIND [2] and GENI [3], and in Japan and Korea the AKARI [4] programs are the main drivers for the definition of the characteristics and capabilities of the Future Internet. In Europe, the Future Internet Research and Experimentation (FIRE) program, the Future Internet

A. Kondoz
Faculty Engineering & Physical Sciences, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, UK
e-mail: a.kondoz@surrey.ac.uk

T. Dagiuklas (✉)
Department of Computer Science, Hellenic Open University, Parodos Aristotelous 18, Patras 26335, Greece
e-mail: dagiuklas@eap.gr

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_1, © Springer Science+Business Media New York 2014

Assembly, and the Future Media 3D Internet Task Force are the main endeavors to provide the fundamental foundations for the Future Internet research [5]. These worldwide initiatives take different approaches to Internet evolution as part of their core objectives but are all related to technological and socioeconomic scenarios as envisioned today.

Future Internet is expected to provide the following characteristics:

1. Internet is experiencing a significant shift from PC-based computing to mobile networks. Mobile data is growing at an unprecedented rate well beyond the capacity of today's 3G, 3G+, and 4G networks. Therefore, mobility has become the key driver for the Future Internet, and convergence demands are increasing across heterogeneous networks.
2. Cognitive radio networks aim to increase spectral efficiency and provide radically new and more efficient wireless access methods.
3. Cooperative communications have recently emerged as strong candidate technologies for many future wireless applications in order to improve performance and throughput of wireless networks, respectively. In particular, theory and experiments have shown that they can be extremely useful for wireless networks with disruptive channel and connectivity conditions.
4. In the last decade, Internet underwent a huge evolution, mainly related to services. The concept of Web 2.0, that is, the second generation of Internet, is led by the generation of new services such as user-generated content, applications, and services, social networks, collaborative webs, and Web TV, together with the growing interactive sets of applications. The new environment has a strong impact on the concept of end-to-end services at all levels, from the applications up to the networks and their underlying technologies.
5. New methods of delivery of content towards the end users. This includes the use of content/information-centric networking architecture making the communication network as location independent possible with the use of distributed architecture (e.g., P2P) associated with streaming, data sharing, routing, and forwarding.
6. Multimedia applications and services make strong demands for cloud computing because of the significant amount in terms of computation resources required to serve millions of both Internet and/or mobile users concurrently. In this cloud-based multimedia-computing paradigm, users are able to store, process, and adapt their multimedia application data in the cloud in a distributed manner. The main drivers of the media-aware cloud architecture are simplicity, efficiency, and scalability. Demands for a media cloud are different from the demands for clouds in other industries.

1.1 3D Media Internet

The 3D Media Internet will play a significant role in achieving the key features of Future Internet [6]. For example, the augmented virtual worlds, the collaborative platforms, and the moving holograms created in 3D Internet will originate new

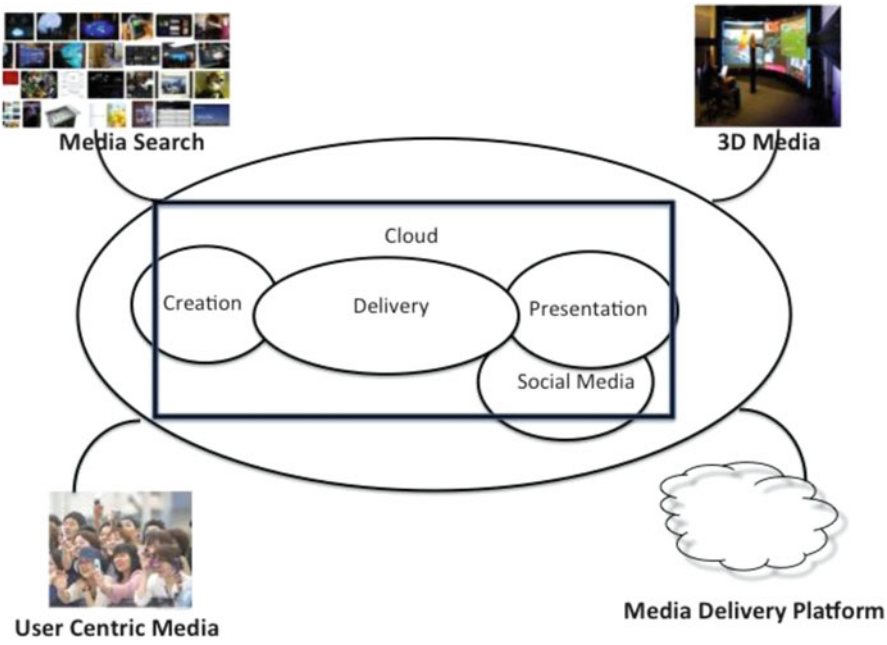


Fig. 1.1 3D media challenges

requirements in terms of information representation and metadata identification. Moreover, the new services and applications will make new demands on ubiquitous user interfaces that will have to support novel inputs (e.g., 3D position sensors), displays (e.g., multiview displays), and presentation (e.g., augmented reality) modalities in different kinds of terminals (e.g., European Commission, 2008).

In these respects, the Future Media Internet will be much more than simply a faster way to go online. It will be designed to overcome current limitations and address emerging trends in areas such as network architecture, content and service mobility, diffusion of heterogeneous nodes and devices, mass digitization, new forms of (3D) user centric/user-generated content provisioning, and emergence of software as a service including interaction with improved security, trustworthiness, and privacy. Figure 1.1 illustrates four 3D Media challenges associated with Media Search, 3D Media, User Centric Media, and Media Delivery Platforms. In order to meet these challenges, the enabling technologies include content creation, delivery technologies, presentation, social media, and cloud computing.

The Future 3D Media Internet has generated a significant amount of recent research goals to overcome current limitations. The resulting innovations could include the following:

1. *3D Content Creation*: In terms of 3D content, the methodological objective is to combine rich interaction-based content production and game-related solutions in order to contribute towards the design and development of novel applications

and services. Instead of the technologically driven approach, content-oriented service development is one of the key requirements for successful Internet applications. A basic understanding and solid background for user behavior can be achieved with the aid of the rich interaction model. Visual elements (look) include texture, light, shadows, reflections, colors, composition, depth perception, shapes, structure, and contrast. Examples of functional concepts (feel) include responsiveness, overall end-user experience, feel of control, how the audiovisual material meets the expectations of the end user, movements, emotional experiences, and affective mechanisms. As opposed to off-line rendering techniques to produce 3D movie effects, in augmented reality applications, the challenge is to perform video processing in real time.

2. *User-Generated Content and Personalization*: Future Internet should provide mechanisms embedded into the network to ease the personalization, adaptation, accessibility, and search aspects but also protect and enforce intellectual property rights related to user-generated content.
3. *Presentation*: In addition it should facilitate a smooth transition from 2D content to 3D content and ease the user participation in 3D content generation and fruition within enhanced 3D collaborative environments. Furthermore, the Future Internet should empower communities to achieve dynamic content creation, provisioning, and sharing (e.g., in social media).
4. *Media Cloud*: Existing cloud computing technologies are not particularly media/video capable. Handling of multiple video flows in terms of encoding, processing, and streaming is a much larger problem that stresses computing infrastructure due to large data and bandwidth requirements. Media cloud infrastructure must be capable of supporting all the functionalities associated with the entire value chain from capturing to processing (encoding, content protection) to content delivery and by greater simplicity to optimize media-related and network infrastructure.
5. *3D Delivery*: P2P technologies have the potential to provide a more cost-effective and flexible delivery solutions for future 3D entertainment services as well as giving users fast interaction with the content and with collaborating partners in their social network. Moreover, recent advances in wireless technologies (e.g., LTE, LTE-A, 5G) will offer the possibility of 3D video to mobile users.
6. *Social Media*: Social networks will play an increasingly important role in the Internet of the future. Internet users will have their own online identity, which will carry them through from one network to the next. Moreover Social Media expects to provide a new type of functionalities and capabilities such as 3D Social Gaming, 3D Life-streaming, and Live-casting.

This book presents recent advances in the area of Future 3D Media Internet. The first four chapters are devoted to the area of 3D Media Coding and presentation. Chapter 2 presents techniques for 3D Media Acquisition and depth map processing. Chapter 3 is a survey of methods for merging the real and the synthetic

in augmented 3D worlds. Finally, Chaps. 4 and 5 present techniques to encode 3D Video and Spatial Audio respectively.

The following chapters cover Networking Aspects for 3D Media. Chapter 6 presents current and future developments in the area of transport protocols for 3D video. Chapter 7 presents application-layer filtering techniques through Media Aware Network Element to optimize 3D Video Delivery. Chapter 8 surveys the techniques used for P2P streaming. Finally, Chap. 9 examines the impact of IP Mobility and mobility management protocols on the quality of 3D video.

The third section presents QoE and QoS advances for 3D Media. Chapter 10 discusses the use of QoS/QoE support in P2P overlays. Chapter 11 makes a survey of QoE methodologies for 3DTV applications and services. Finally, Chap. 12 presents error concealment strategies and techniques to improve QoE in multiview video applications.

The last section presents two 3D applications. Chapter 13 describes the use of 3D images and video in medical surgery and training applications to improve diagnosis and surgery operation. Finally Chap. 14 presents innovations in 3DTV capturing, data representation, compression, transmission, and rendering required for a technically efficient and commercially successful 3DTV broadcast system.

References

1. http://www.ict-fire.eu
2. http://www.nets-find.net/
3. http://www.geni.net/
4. http://en.wikipedia.org/wiki/AKARI_Project
5. Tselentis, Georgios (ed) (2009) *Towards the Future Internet: A European Research Perspective*. IOS press
6. Zahariadis, Theodore, Petros Daras, Isidro Laso-Ballesteros (2008) *Towards future 3d media internet*. *NEM Summit*:13–15

Part I
3D Media Coding and Presentation

Chapter 2
3D Media Representation and Coding

Pedro Assunção, Luís Pinto, and Sérgio Faria

Abstract Nowadays, three-dimensional (3D) multimedia provides immersive user experiences in virtual and real 3D environments, based on advanced technology that is rapidly evolving towards different fields of application. An important element in 3D multimedia is undoubtedly the visual information, where 3D video plays a major role. This chapter addresses such element of 3D multimedia, by presenting a comprehensive description of the most common 3D video formats used in various 3D multimedia services and applications. Uncompressed 3D video representation is described, specifically focusing frame compatible formats used for backward compatibility with 2D systems, followed by those formats that explicitly include depth information, either in single-view or multiview representations. Then an overview of standard coding algorithms, currently used for 3D video coding, is presented along with a discussion of their main characteristics in terms of processing methods and performance. Since the response of the human perceptual system to 3D visual content includes specific features different from those already known from 2D perception, the last sections of the chapter describe recent studies dealing with asymmetric representation and coding of stereoscopic video. The use and adaptation of standard coding methods to benefit from asymmetric characteristics of the human visual system is presented and discussed in the light of recent research advances. The chapter concludes by highlighting the most relevant issues in the current context of 3D video representation and coding.

2.1 Introduction

The human perception of the real world is three dimensional (3D) and necessarily includes information about volume and depth, which is not explicitly included in classic representations of natural scenes using digital multimedia signals. The huge

P. Assunção (✉) • L. Pinto • S. Faria
Instituto de Telecomunicações/Instituto Politécnico de Leiria, Delegação de Leiria, Portugal
e-mail: paassunc@ieee.org

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_2, © Springer Science+Business Media New York 2014

amount of structured data required to represent 3D visual scenes leads to the need of standard representation formats in both uncompressed and compressed domains. Since 3D video is the most common type of media content used to provide 3D immersion experience, this is also a driving factor of technology evolution in several domains, ranging from high-resolution 3D cinema and 3D television to small screen applications (e.g., games) using autostereoscopic displays.

Recent evolution of 3D media services along with increasing penetration of 3D-ready equipment in the consumer market leads to the coexistence of emerging 3D systems with legacy ones. Several 3D representation formats are currently used to enable efficient coding for storage and transmission across interoperable systems also enabling operation with equipment in different technological evolution stages either in the segment of professional or consumer market.

The most common 3D video formats are described in the next sections, focusing frame compatible formats used for backward compatibility with 2D systems followed by specific formats explicitly including depth information either in single-view or multiview representations. Then an overview of standard coding algorithms currently used for 3D video is presented along with their main characteristics in terms of processing architectures and coding performance. Since the characteristics of the human perceptual system regarding 3D visual content include specific features, different from those known from 2D perception, the last sections of this chapter describe recent studies dealing with asymmetric representation and coding of stereoscopic video. The use and adaptation of standard systems to benefit from asymmetric characteristics of the human visual system is presented and discussed in the light of recent research advances.

2.2 Raw Formats for 3D Video

There are several different formats used to represent raw 3D video. The common requirement is that all of them must provide means for stereoscopic viewing, since this is the underlying principle behind depth perception from visual information. In the following sections, different formats are described including those based on stereo views, 2D views plus explicit depth information, and multiple views also associated with depth maps.

2.2.1 Frame Compatible Formats

In the context of 3D multimedia services and applications, 3D video representation through frame compatible formats is a key factor to guarantee compatibility with existing 2D video networking technology and equipment. Successful deployment of 3D video delivery services and applications is enabled by making possible transmission of 2D-compatible formats over current networks and legacy decoders with

Table 2.1 Standard frame compatible formats

ID	Compatible format
0	Checkerboard
1	Column-based interleaving
2	Row-based interleaving
3	Side-by-side
4	Top-bottom
5	Temporal interleaving
6	2D
7	Tile format

3D-ready displays already common in the user market. By using frame compatible formats, seamless compression of 3D video content is also accomplished with existing 2D encoders without the need to modify the coding algorithms.

In the case of stereoscopic video, representation in 2D-compatible formats is achieved by multiplexing the two different views into a temporal sequence of 2D signals. This means merging two different views into a classic sequence of single-frame representation. If the full resolution of the two views is maintained, then such representation format has twice the resolution of its equivalent 2D. However, taking into account that good perceived quality of 3D video does not necessarily require two high-quality views, either one or both views may be subsampled in one dimension in order to fit two high-definition (HD) frames into only one HD frame slot [1]. Identification of left and right views is done via specific signaling, used to distinguish the data representing each one. Using H.264/AVC to encode 3D frame compatible formats, the recommended signaling method is the use of supplemental enhancement information (SEI) messages, as shown in Table 2.1, where the *frame_packing_arrangement_type* field of the SEI message is defined according to subclause D.2.25 in the standard [2]. SEI messages are used to convey information about how decoders should handle their output according to the frame-packaging scheme used. There is also an SEI value defining 2D format, which enables switching between stereo and non-stereo content. Additionally to the type of content, the SEI message includes other fields such as the arrangement *id* that can be used to identify which frame is the left or right view.

In the following subsections, these frame compatible formats are described along with their main characteristics.

2.2.1.1 Side-by-Side

The side-by-side format is shown in Fig. 2.1a. In this format, the two stereoscopic views are concatenated side by side, giving rise to a single 2D matrix of pixels with the same resolution in the vertical direction while in the horizontal one there is twice the number of pixels of each single view.

However, since doubling the spatial resolution has strong implications in compressed rates, the horizontal resolution of the original views might be halved

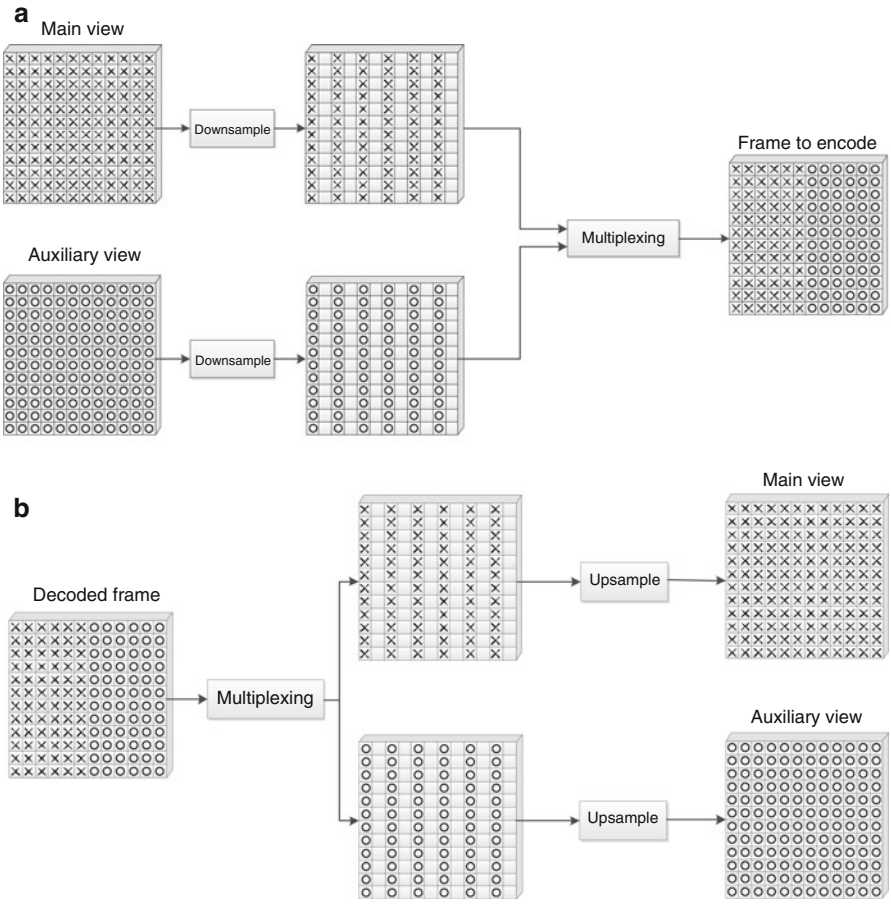


Fig. 2.1 (**a**) Side-by-side packing arrangement. (**b**) Side-by-side up-conversion process

through downsampling before packing into this side-by-side arrangement for subsequent encoding and transmission. Correspondingly, the counterpart up-conversion process must be done after decoding to display the full-resolution stereo images, as shown in Fig. 2.1b.

A different version of the side-by-side arrangement can be accomplished by sampling the stereoscopic views using a quincunx pattern. In this case, even though the horizontal resolution is also reduced to half of the original, still half of each view columns is maintained, as shown in Fig. 2.2. Quincunx sampling relates better with the characteristics of the human visual system (HVS) in terms of frequency domain representation. Thus, this sampling pattern preserves more relevant information from the original signal, which has the potential to result in better perceived quality.

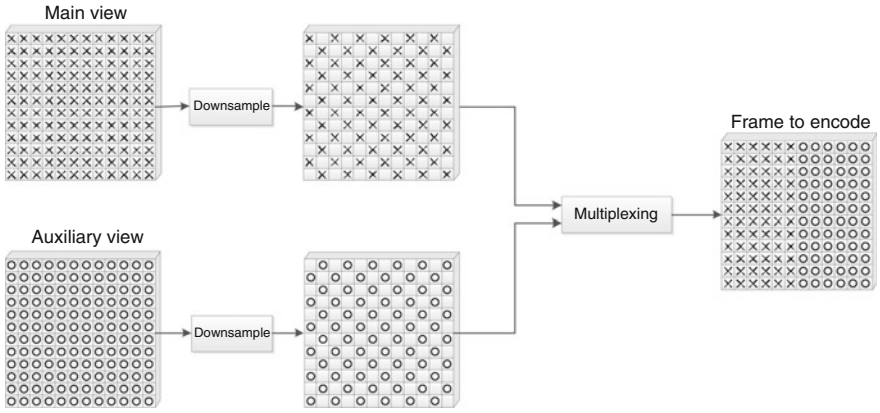


Fig. 2.2 Side-by-side with quincunx arrangement

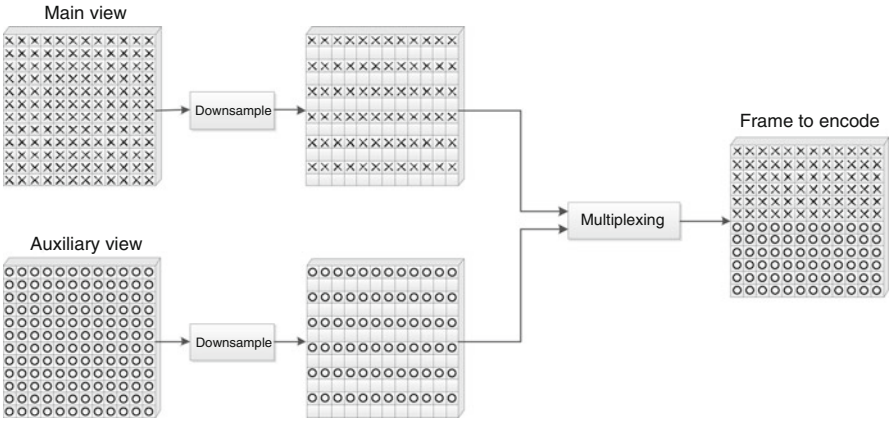


Fig. 2.3 Top-bottom arrangement

2.2.1.2 Top-Bottom

The top-bottom format is based on the same concept as side-by-side, but in this case downsampling is applied to the vertical resolution and the resulting frames concatenated as shown in Fig. 2.3.

Unless otherwise specified by the SEI message, in standard top-bottom format the downsampled left view is concatenated into the first half (top) of a composite frame while the downsampled right view is concatenated into the bottom half. This 3D frame compatible format should not be used with interlaced source material because the vertical resolution of interlaced fields is already half of the full-resolution frame and further downsampling in this dimension could incur in too much loss of spatial detail.

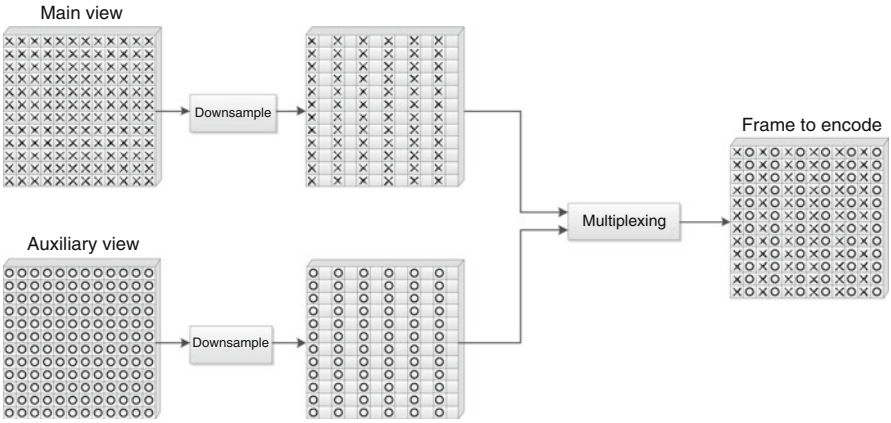


Fig. 2.4 Column interleaving arrangement

Both side-by-side and top-bottom formats are preferred for production and distribution of 3D content in comparison with the ones described next, based on spatial interleaving. This is because interleaving is prone to cross-talk artifacts and color bleeding.

2.2.1.3 Interleaved Formats

Interleaving methods provide higher correlation in the composite frame by multiplexing both downsampled views, either vertically or horizontally according to the downsampled dimension. If both views are half sampled in the horizontal dimension, then a column interleaving arrangement is reached, as shown in Fig. 2.4. Otherwise, if downsampling is performed in the vertical dimension, then multiplexing is done row by row, attaining a row interleaving arrangement for the composite frame.

These interleaving methods can be further combined in order to create a multiplexed frame-like checkerboard such as the one depicted in Fig. 2.5. In this type of format, each view must be downsampled using checkerboard nonmatching patterns. In the case of the left view, this means that in odd rows each other pixel should be kept starting from odd columns, while in even rows each other pixel should kept, starting from even columns. In the case of the right view the complementary pattern must be used.

Other frame compatible arrangement is based on interleaving in the temporal dimension. In this type of interleaving the frame rate of each view is reduced to half of its original rate and then the even frames from the left view are temporally multiplexed with the odd frames from the right view, as shown in Fig. 2.6. In this type of format the spatial resolution of the original views is maintained. It can be particularly suitable to represent low motion 3D content, where frame rate is not a very relevant requirement.

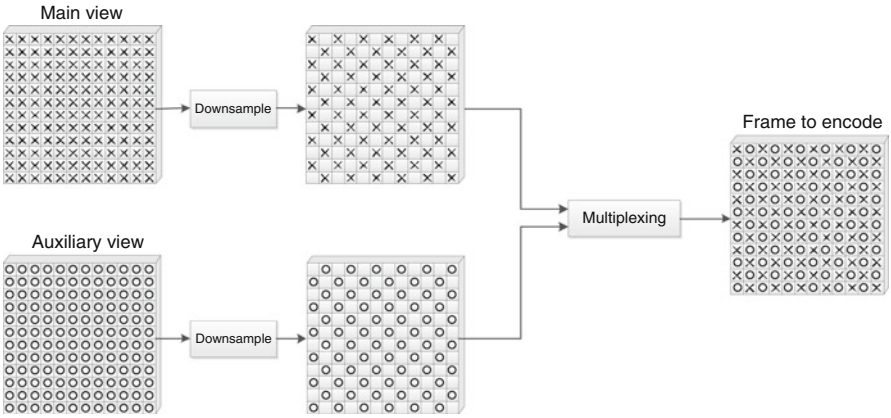


Fig. 2.5 Checkerboard arrangement format

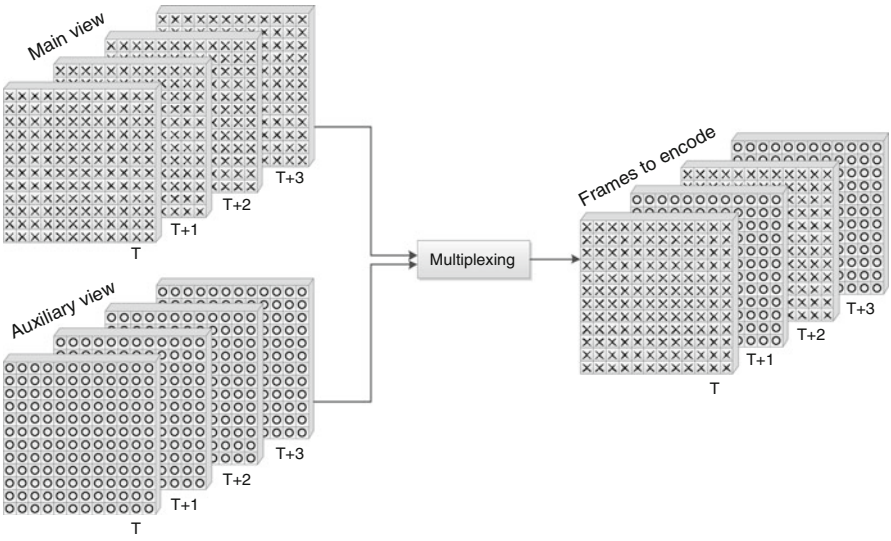


Fig. 2.6 Temporal interleaving frame arrangement

2.2.1.4 Tile Format

The last amendment of H.264/AVC in regard to the use of frame compatible formats introduced the tile format [3, 4]. The arrangement depicted in Fig. 2.7 allows two HD frames (1,280 per 720 pixels) to be packed into a Full HD frame

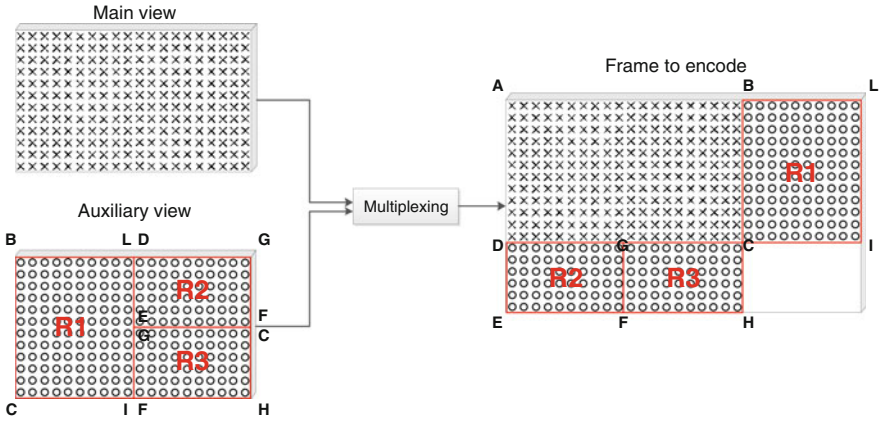


Fig. 2.7 Tile frame arrangement

(1,920 per 1,080 pixels) using a tiling method, where different regions of one view are tiled with the other view.

As seen in the figure, the left view is located at the top left corner of the Full HD frame without any type of preprocessing. The right view is then divided into three regions (R1, R2, and R3), which are placed in specific regions of the resulting Full HD frame.

The great advantage of this method is the backward compatibility with legacy 2D devices, as it requires only a 720p crop to obtain a 2D version of the content in HD resolution. Moreover there are no downsampling operations involved, which means that full resolution of the original frames is maintained in all dimensions.

A potential drawback introduced by this method is lower coding efficiency and annoying artifacts in the coded images due artificial edges created by tiling. However objective tests conducted show that these are not problematic as impairments are not noticeable (comparing with simulcast), mainly above 1 Mbps [4].

2.2.2 Video Plus Depth

An alternative to stereoscopic representation of 3D video consists in two separate 2D signals to convey color image information and the depth associated to each pixel, i.e., the distance to the camera [5]. Such information being available at display side can enable the generation of virtual views through depth-based image rendering techniques. Known as video plus depth ($V + D$), this format has implicit higher complexity than stereo views because it requires either additional computation to obtain the depth values from multiple views of the scene or specific image acquisition hardware to obtain the depth maps directly from the scene, e.g., using hybrid camera systems [6].

Fig. 2.8 Sequence breakdance: video (*left*) plus depth (*right*)

Depth values are usually represented as integers in the range of 0–255, thus using 8 bits per pixel which results in a gray-scale depth map, as shown in Fig. 2.8. These values translate to the maximum and minimum distance of each point. A warping function should be used to reconstruct a stereoscopic sequence by synthesizing the other view of a stereo pair from the color image and the corresponding depth map. Depth image-based rendering (DIBR) is a method commonly used for this purpose [7]. In view synthesis using DIBR there are some problems that may result in image distortions, such as the possibility of occlusions due to the lack of unique texture data that may be needed to render the other stereo view through the depth map.

Separate encoding of each signal (video and depth) is possible by a standard monoscopic codec, such as H.264/AVC. In regard to encoding the depth map, the gray-scale values can be given to the encoder as the luminance component of pseudo video signal where chrominances are set to a constant value. Since the color video is encoded as regular monocular video, this format has inherent backward compatibility with legacy decoders. This format allows extended possibilities at the receiving side compared to the traditional stereo video. For instance, it is possible to adjust the amount of depth perceived by viewers by adjusting view synthesis or to render several different virtual views for multiview displays.

2.2.3 Multiview Video Plus Depth

As mentioned before, the $V + D$ format is particularly suited to multiview displays because it enables generation of different virtual views of the same scene. However, if a wide range of views is required, the $V + D$ format is no longer suitable because not many different views can be rendered with enough quality from only one view and corresponding depth map. This is because the original view may become farther away than the one to be synthesized producing visible artifacts due to occlusions, which cannot be properly handled in such cases. This is mainly relevant in wide-range multiview (autostereoscopic) displays or free viewpoint video applications.

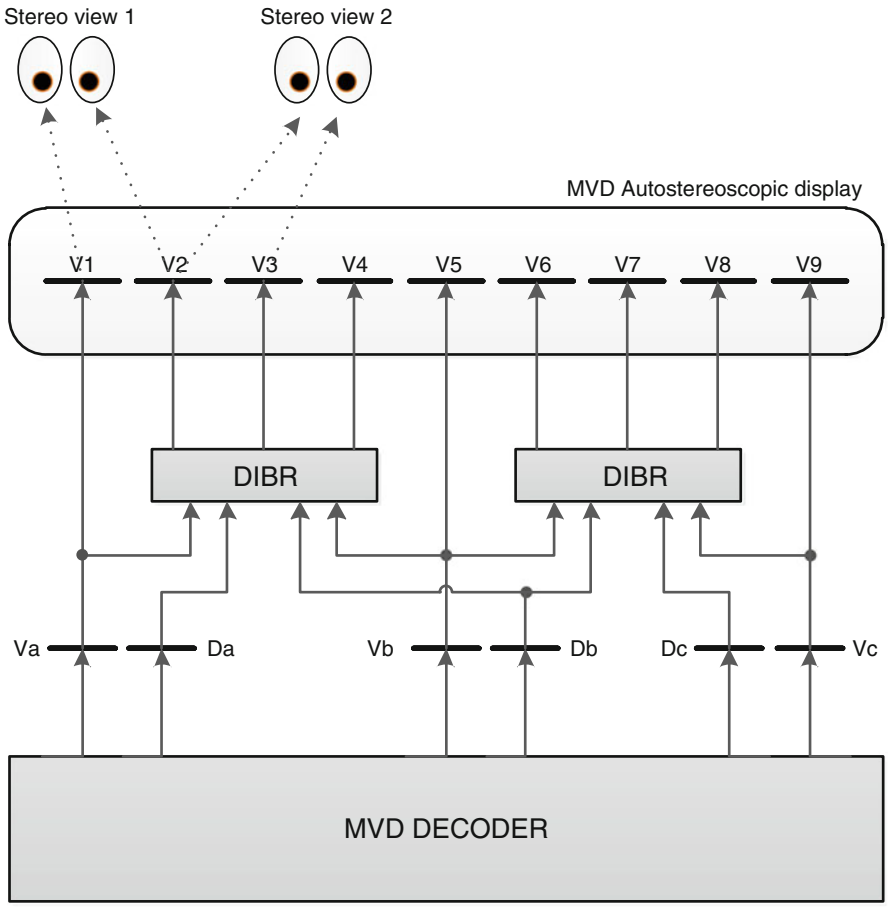


Fig. 2.9 Use of MVD format

The multiview video-plus-depth (MVD) format provides a solution for generating many virtual views by including several views from the same scene, each one with an associated depth map. Based on each pair (V_n, D_n), $n = 1, \ldots N$, it is possible to render virtually any intermediate view, giving rise to free viewpoint video. Figure 2.9 shows an example with an autostereoscopic display, where 9 views are generated from only 3 views plus their associated depth maps. These 3 views are actually the only ones available to render the other 6 virtual views [8].

MVD has similar complexity issues as $V + D$ since it also requires depth acquisition/estimation at the sender side and rendering stereo views at the decoder. On the one hand, this format allows significant savings in storage and transmission requirements, but on the other hand higher processing complexity is required either in the display side only or in both sides, i.e., acquisition and display.

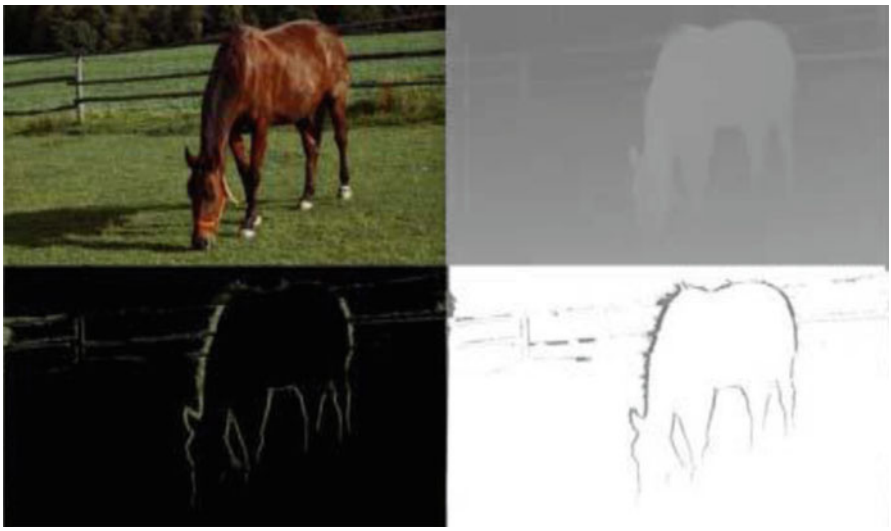


Fig. 2.10 Example of LDV

2.2.4 Layered Depth Video

The layered depth video (LVD) is another 3D video format comprising color images, depth maps, and an additional layer providing occlusion information, which is used to fill up occluded regions in the rendering process. Such additional layer contains texture information from those regions in the scene that are occluded by foreground objects in the available view. Using this format, rendering of virtual views benefits from layered information because it includes the necessary data to synthesize virtual views, which otherwise would be missing. In LDV it is also possible to represent residual layers to include visual data that is not available in the main view but visible from other viewing directions. Figure 2.10 shows an example of a color image and its depth map (top) along with an occlusion layer and depth map (bottom).

A variant of LDV is known as depth-enhanced stereo (DES), which basically includes two sets of LDV data [8]. Since LDV is an extension of MVD, its inherent computational complexity is higher than MDV due to the operations (warping, matrix subtraction/addition, etc.) that are necessary to obtain the residual layers. On the receiver side, rendering the extra views using LDV has similar computational complexity as MDV. Besides the ability to better cope with occlusions, LDV has also the advantage of requiring a smaller amount of data than MVD for multiview 3D video representation. However, since LDV relies on residual data, the potential impact of transmission errors or data loss is greater than in MDV.

2.3 3D Video Coding

Current 3D video technology is mostly based on stereo systems, but this is rapidly evolving to include more information, that may be either more views and/or depth information. These additional data is used to feed 3D systems that are able to provide richer immersive experiences to users. Regardless of the 3D video format used in such systems, these data have to be encoded, in order to fulfill the application and service requirements and to achieve useful compression ratios. Hence, due to the multidimensional nature of this content, it can be either jointly encoded as a whole, by exploiting their correlation, or separately as a set of independent sources.

2.3.1 Simulcast

Simulcast refers to independent encoding and transmission of several views and possibly their corresponding depth maps, using any encoder to encode each data sequence. To simulcast stereo or multiview video, each view (left, right) is independently encoded without using any type of inter-view prediction, as can be seen in Fig. 2.11. Any standard video encoder such as H.264/AVC [2] or HEVC [9] can be used for this purpose. In this case, the complexity and processing delay is kept low, since dependencies between views are not exploited, and backward compatibility with 2D systems is maintained by decoding only one view.

However, a simulcast solution has a drawback in the coding efficiency, as it does not exploit the inter-view redundancy. In this sense, studies on asymmetric video coding suggest that one view may be encoded with less quality than the other, with significant bit rate savings. Such scheme may be implemented by means of coarse quantization or by reducing the spatial resolution [10]. This can be achieved without loss of the stereo perception, but when played for long periods, the unequal quality for each eye may cause eye fatigue. To overcome such effect, the toggling of quality between both views has been suggested [11]. A more detailed description of asymmetric coding is presented in Sect. 4 of this chapter.

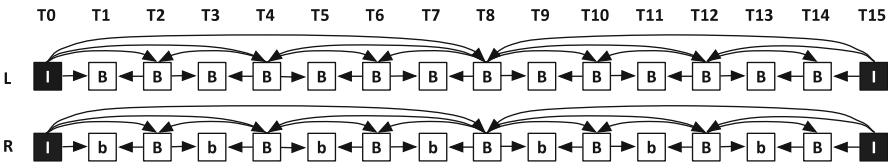


Fig. 2.11 Simulcast structure

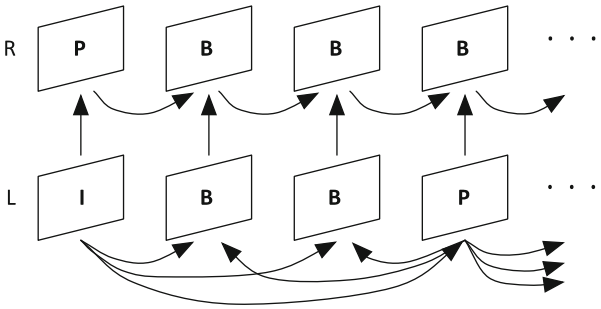


Fig. 2.12 Prediction structure in stereo video

2.3.2 Multiview Video Coding

Multiview video coding (MVC) comprises joint coding of two or more views of the same scene. When only two views are allowed, this is named as stereo video coding. Figure 2.12 shows an example of a stereo video sequence and frame coding dependencies regarding the right (R) and left (L) views. The left view is independently encoded to ensure compatibility with 2D video systems, while the right view uses inter-view prediction from the left one achieving higher coding efficiency at the cost of greater coding complexity and dependency.

The first standard for multiview applications was the extension of the MPEG-2 MVP [13] (Multiview Profile), which has been approved as International Standard in 1996 when it was envisioned to be a profile appropriate for applications requiring multiple viewpoints. The underlying coding principles used in this codec are mostly the same as those currently used in more advanced ones. The general architecture is depicted in Fig. 2.13, where the base layer (left view) is encoded as monoscopic video to maintain compatibility with the MPEG-2 video decoders at the Main Profile.

As shown in the functional diagram of Fig. 2.13, the enhancement layer (right view) is encoded using hybrid prediction of motion and disparity and temporal scalability tools. By exploiting the similarity between the left and right views, higher compression of the right view was achieved. Both layers have the same spatial resolution at the same frame rate. For MVP, an extension has been introduced specifying the height of image device, the focal length, the F-number, the vertical angle of the field of view, the position and the direction of the camera, and upper direction of the camera.

A further step towards multiview compression has been done in 2009 with an amendment of H.264/AVC (Annex H) to support MVC with the *Multiview High Profile* [2]. A typical frame structure and interframe/view coding dependency is illustrated in Fig. 2.14. Two types of predictions are explicitly used to achieve increased coding efficiency: intra-view and inter-view prediction. The prediction structure determines the decoding sequence according to the dependencies between frames. Depending on the acquisition arrangement, any two adjacent views may comprise a stereo pair to provide 3D viewing experience.

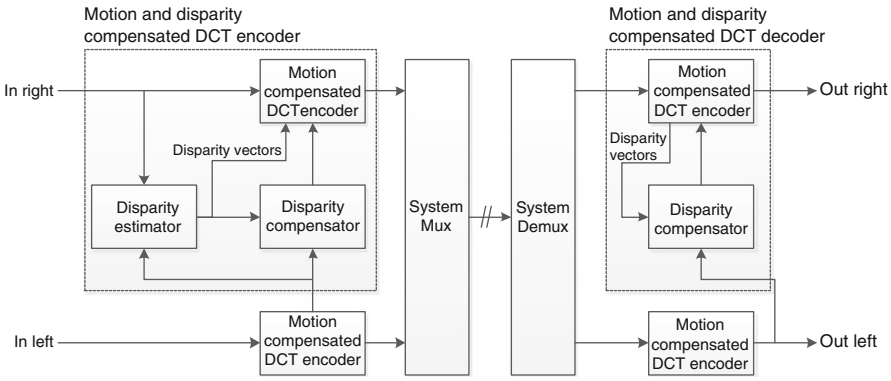


Fig. 2.13 Codec reference model for the MVP [12]

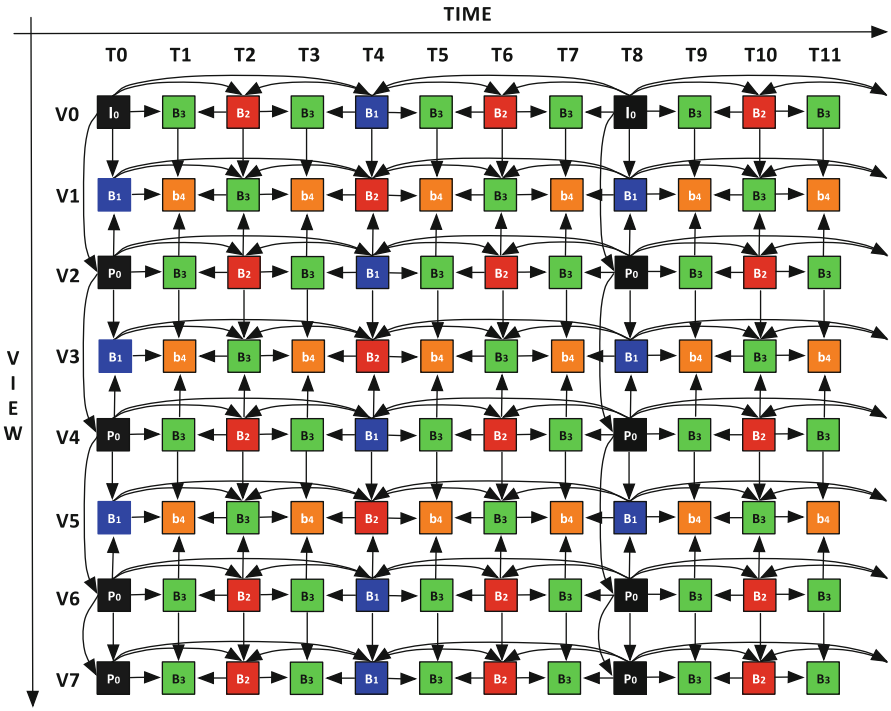


Fig. 2.14 Typical MVC frame coding structure

The standard extension for MVC supports a flexible reference picture management that is used by the inter-view prediction scheme, i.e., the decoded frames from other views are made available for prediction in the reference picture lists. This scheme allows a particular view to have some blocks predicted from temporal

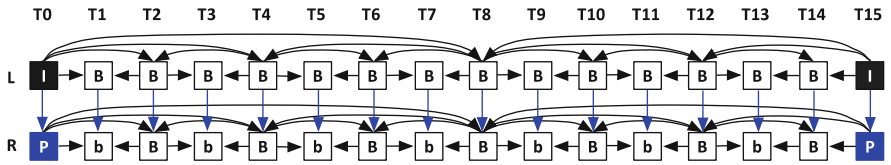


Fig. 2.15 MVC Stereo High Profile frame coding structure

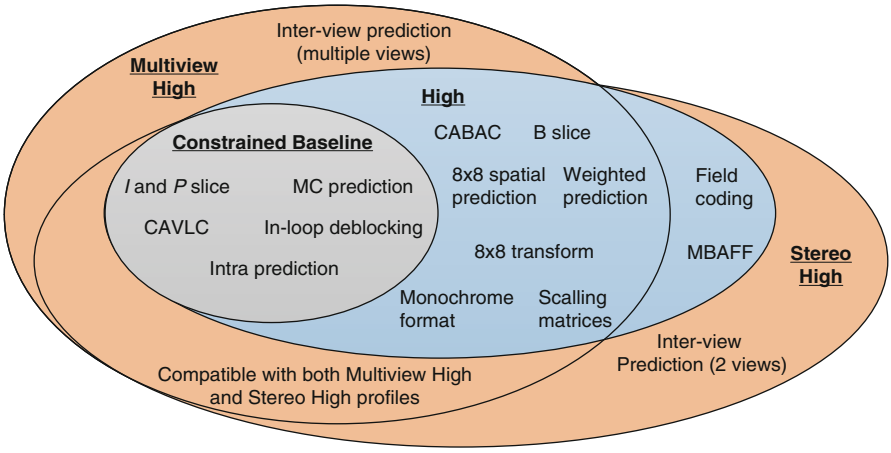


Fig. 2.16 MVC profiles and tools associated

references while others can be predicted from inter-view references. MVC includes a mandatory base view in the compressed multiview H.264/AVC stream, which can be independently extracted and decoded for 2D viewing. To decode other views, information about view dependency is required. As in previous schemes, unequal rate allocation may be used across the views.

In 2010, H.264/AVC was extended with the *Stereo High Profile* for stereo video coding, with support for interlaced coding tools. In this case, the frame coding structure only stands for two views, as can be seen in Fig. 2.15.

Figure 2.16 shows the two MVC profiles, *Multiview High* and *Stereo High*, highlighting the fact that a common set of coding tools is used in both of them. When coding a video sequence with two views using noninterlaced tools only, the coded stream conforms to both profiles.

2.3.2.1 Coding Performance

A relevant performance metric of stereo video encoding is the amount of additional bit rate required to encode a second view using standard encoders. The authors in [15] carried out a performance study using MVC High Profile to encode 9

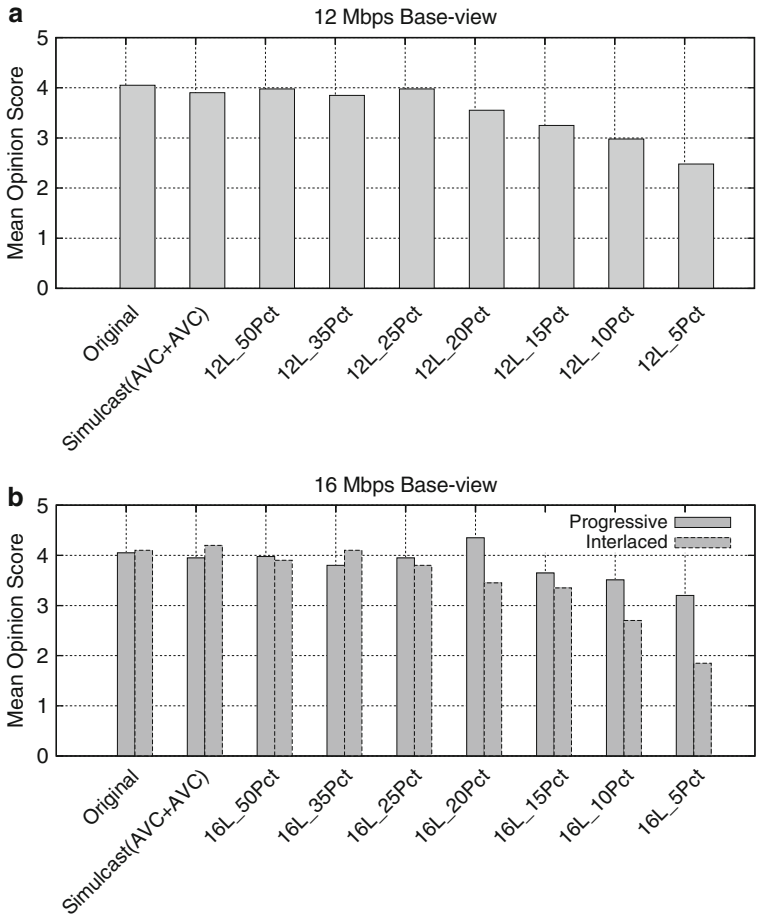


Fig. 2.17 Subjective quality evaluation MOS (mean opinion score) [15]

high-definition (HD) 3D video clips with various types of content, from animation and live action shots, including progressive and interlaced material. The base view was encoded at 12 Mbps and 16 Mbps. The dependent view was encoded at a wide range of bit rates, from 5 to 50 % of the base-view rate, achieving combined bit rates from 12.6 to 24 Mbps. The subjective results were rated by viewers based on a numeric value, where 5 means excellent and 1 very poor. Figure 2.17 shows the relevant results.

The base view encoded at 12 Mbps plus the second view at N % overhead is labeled as 12L_NPct, where N corresponds to Mbps. As can be observed in the case of 12Mbps in Fig. 2.17a, the results of subjective quality evaluation suggest that, as compared to 2D, 20–25 % bit rate increase would provide satisfactory picture quality in 3D HD video applications. These results show that it is possible to trade off between the bit rates of the base and second views for given a total

bandwidth. Higher bit rate for the dependent view preserve better the 3D effect, but in this case fewer bits are left for the base view. For instance, for a bandwidth of 18 Mbps, both 12L_50Pct and 16L_15Pct (Fig. 2.17b) can be chosen. However, it may be more important to preserve the picture quality of the base view, in order to guarantee good quality 2D viewing.

The performance of the MVC encoder using inter-view prediction tools, against simulcast, has also been tested over a broad range of test material, using the common test conditions and test sequences specified in [16]. The results, given as Bjontegaard Delta measurements (BD bit rate and BD PSNR) [17], demonstrate that MVC with up to 8 views can save on average 24 % of bit rate (1.2 dB in BD PSNR gain) in comparison with the total simulcast bit rate, at the same equal quality in each view [14].

The MVC Stereo High Profile also supports stereo coding for interlaced sequences. An experimental performance comparison between simulcast and MVC is presented in [18]. The LG's stereo 1080p24 HD sequences and other stereo sequences, obtained by selecting two views from the multiview MVC test sequences, were used with coding patterns for interlaced defined as IBBP and hierarchical B frames for progressive video. It was found that, for progressive stereo sequences, MVC achieved 9.36 % and 11.8 % of average bit rate savings for both views, respectively, for the stereo MVC test sequences and LG's stereo 1080p24 HD progressive sequences. For interlaced video sequences, an average of 6.74 % bit rate saving was achieved, for both views (two SD and one HD interlaced sequence). The estimated savings for the dependent view is twice the total saving, that is, approximately 20 % gain in the progressive scan sequences and 15 % in the interlaced scan sequences.

The coding performance of the MVC Stereo High Profile was also compared against AVC High profile (Simulcast), using 8 stereo sequences of 3D cinema clips (1,920 $\times$ 1,080 at 24 fps with a playback time of 20 s for each view), also with contents from animation to live shots. The coding pattern IBBP was used along with the same fixed QP (quantization parameter) matrix for both views and an intra (I) period of 1 s.

The Stereo High Profile achieved an average coding efficiency gain of about 15 %, in both views, as compared to simulcast using AVC High profile. For some Hollywood 3D cinema clips, the coding efficiency gain can go up to 21 % [19]. Regarding the right view only, as can be seen in Fig. 2.18, the BD bit rate gain for *Clip 01* is 42.17 % (or 2.00 dB in BD PSNR) and for all 8 clips the BD bit rate gain is 30.50 % (1.07 dB in BD PSNR), on average. Concerning the coding efficiency gain, this is more pronounced in the animation sequences. The content of live shots very often presents different sharpness, brightness, color, contrast, etc., in the right view, as compared to the left view. This difference reduces the efficiency of the inter-view prediction.

The recently approved standard for high efficiency video coding (H.265/HEVC) [9] has emerged with a new set of coding tools, which have significantly increased the compression efficiency in comparison to the previous video compression standards. Thus, quality evaluation tests demonstrate that HEVC is able to

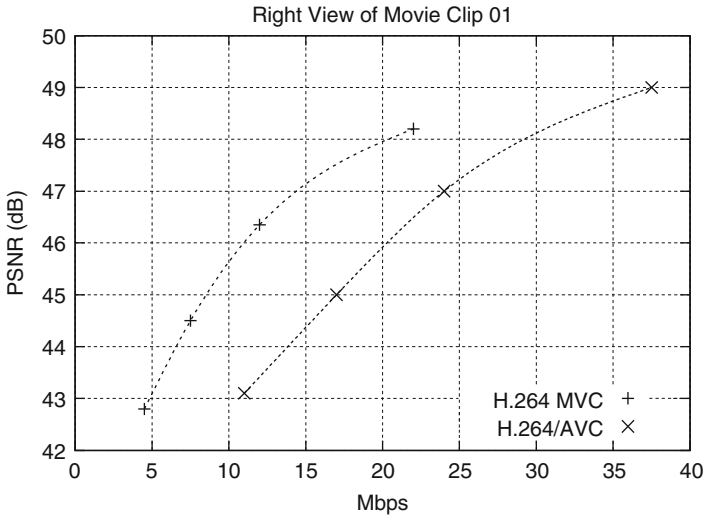


Fig. 2.18 Sample coding results for *Clip 01* stereo sequence [19]

increase the compression ration about 50 % at the same quality, in comparison with H.264/AVC. This has also been reported upon subjective tests [20]. However, such increased efficiency is obtained at the cost of higher computational complexity, as HEVC requires 2–10 times more computation in the encoder when compared with H.264/AVC. At the decoder side, HEVC presents similar complexity to that of the H.264/AVC [21].

The extension of HEVC to multiview video coding (MV-HEVC) is described in Annex F of Recommendation ITU-TH.265, supporting 3D applications such as stereoscopic television. This extension will enable HEVC-based high-quality 3D video coding, at approximately half of the bit rate required by previous services and applications like 3D television and 3D Blu-ray discs. Similarly to the MVC extension of H.264/AVC, MV-HEVC takes advantage of the inter-view prediction tools to exploit the redundancy between views.

Due to its powerful encoding tools, simulcast with HEVC, i.e., each view independently encoded, outperforms H.264/MVC. This is clearly observed in Fig. 2.19 where rate-distortion performance is compared. In the case of H.264, where MVC provides significant bit rate reduction when compared to AVC simulcast, MV-HEVC also outperforms HEVC simulcast [22]. Regarding multiview, MV-HEVC halves the bit rate required by MVC to encode a multiview sequence, in the same proportion as HEVC improves the coding efficiency when compared to H.264/AVC coding of a single-view video. Similarly to MVC, MV-HEVC provides backward compatibility to allow single-view decoding. In MV-HEVC, inter-view prediction allows the inclusion of inter-view reference pictures in the reference picture lists that are used for prediction. This is also similar to H.264/MVC.

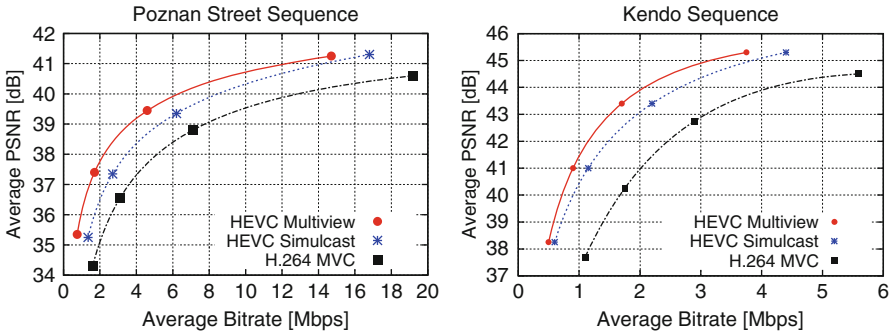


Fig. 2.19 R-D performance for "Poznan_Street" (*left*) and "Kendo" (*right*) for 3 views [22]

2.3.3 Video Plus Depth Coding

As previously described, depth-based representations are emerging as an important class of 3D formats, enabling the generation of virtual views through DBIR techniques. Thereby, this format enables display-independent solutions, as different displays may generate more views as required. Although the depth data is not directly output to a display and viewed, maintaining the fidelity of depth information is very important while encoding because it has great influence in the view synthesis quality, due to the geometric information provided by depth. Thus, reaching a good balance between compression ratio and quality of coded depth data is of utmost importance. Note that depth information can be used by encoders to attain more efficient compression, through view synthesis prediction schemes.

The ISO/IEC 23002–3 standard, also referred to as MPEG-C Part 3, specifies the representation of auxiliary video and supplemental information [23]. This is the first standard where signaling of coded depth map is explicitly allowed to support the coded format of video-plus-depth. It is worthwhile to notice that this standard does not specify the coding standard that should be used for depth and video information, which allows compatibility with any legacy receiver. The use of MPEG-C Part 3 is illustrated in Fig. 2.20a with two H.264/AVC encoders generating two streams (BS), one for video and another for depth. Then these streams are multiplexed using frame-by-frame interleaving before encapsulation into a single transport stream (TS). At the receiving side, these streams are demultiplexed and independently decoded to produce the output video and depth signals, which in turn are used to generate the second stereo view.

Another possibility for encoding and transmission of video plus depth is to use the "Auxiliary Picture Syntax" defined in the H.264/AVC standard, which defines that auxiliary monochrome pictures can be sent with the video stream, i.e., the primary coded pictures may be associated with other types of data, jointly encoded but not used for display. This is illustrated in Fig. 2.20b where depth is treated as the auxiliary picture of the color view. The H.264/AVC encoder is given both

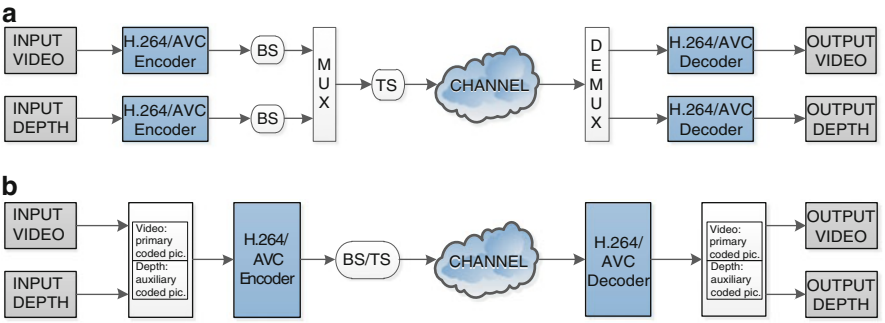


Fig. 2.20 $V + D$ coding and transmission system

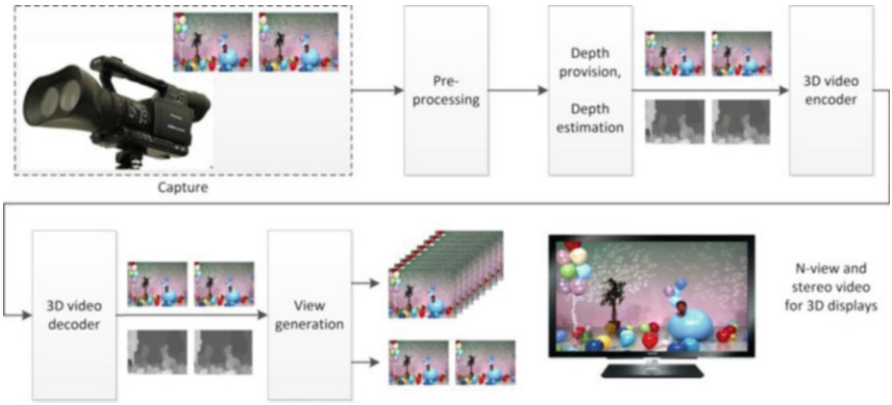


Fig. 2.21 Multiview plus depth coding and transmission

sequences to be jointly encoded, producing one single coded stream (BS/TS). The color images and corresponding depth maps are seamlessly decoded as single pictures and then separated into two different signals for viewing and synthesis of the second stereo view.

As mentioned before, in order to overcome the drawbacks of the 2D plus depth format in generating diverse virtual views, the MVD format allows enhancing the 3D rendering capabilities at a reduced number of transmitted views plus corresponding depth.

In Fig. 2.21, a general coding and transmission system for MVD is depicted. A few cameras (2 in the case of Fig. 2.21) are necessary to acquire multiple views of the scene while the depth information can be estimated from the video signal itself by solving for stereo correspondences or directly provided by special range cameras. Depth may also be an inherent part of the content, such as with computer-generated imagery. Both types of signals can be either independently or jointly encoded using any coding schemed previously described. At the receiver, the few decoded views and their corresponding depth maps are used to generate a

higher number of virtual views as necessary for each particular multiview service or application.

Since depth information mainly consists of larger homogeneous areas and sharp transitions along object boundaries, the frequency spectrum of a depth map mostly comprises low and very high frequencies. As a depth sample represents a spatial shift in the color samples of the original views, coding errors result in wrong pixel shifts in synthesized views, which may lead to annoying artifacts, especially along object boundaries.

Joint video and depth coding is the current path for finding a scheme able to achieve high compression efficiency. For instance, it has been shown that coding efficiency can be significantly increased using scene geometry information such as depth maps [5]. Beyond the use of inter-view prediction techniques, which can be applied to both video and depth independently, there is some block-level information, such as motion vectors, that may be similar for both data types and thus can be shared. In the context of multiview coding, adjacent views may be warped towards reference views in order to reduce the residual error between views. However, since video compression algorithms are typically designed to preserve low frequencies, to maintain the fidelity of edge information in depth maps special coding techniques are required to deal with such particular characteristics.

To evaluate the compression efficiency achieved by the new coding tools implemented in HEVC, several performance studies have been carried out [24]. Different schemes have been compared, namely, HEVC simulcast (based on HM 6.0), MV-HEVC (multiview HEVC), and 3D-HEVC (both based on HTM 3.1). MV-HEVC is a simple extension of HEVC, using the same principles of H.264/MVC framework, providing backward compatibility for 2D video decoding and utilizing inter-view prediction. The 3D-HTM encoder is an extension of HEVC, where the base view is fully compatible with HEVC and the dependent views use additional tools to exploit motion correlation and mode parameters between base and dependent views. Using the common test conditions for 3D video coding (Doc. JCT3V-A1100), for an average of 7 sequences, HTM is able to achieve gains up to 22,7 % over MV-HEVC and 47,6 % over HEVC simulcast, measured as Bjøntegaard delta bit rates.

The 3DV standardization is expected to be finalized by early 2014, which probably include 3D AVC (based on H.264/AVC) and 3D-HTM (based on HEVC). These schemes allow the choice of MVD, reducing the number of transmitted views and enabling joint encoding of view depth. It is also possible that single depth and asymmetric frame sizes for view and depth will be supported.

2.3.3.1 Depth Coding

Extensive research has been done on efficient coding algorithms for depth as well as in the use of depth for video coding. Correlation between depth maps from different views has been exploited and decoded information from texture components was found useful for depth decoding, e.g., the motion prediction modes. In the case

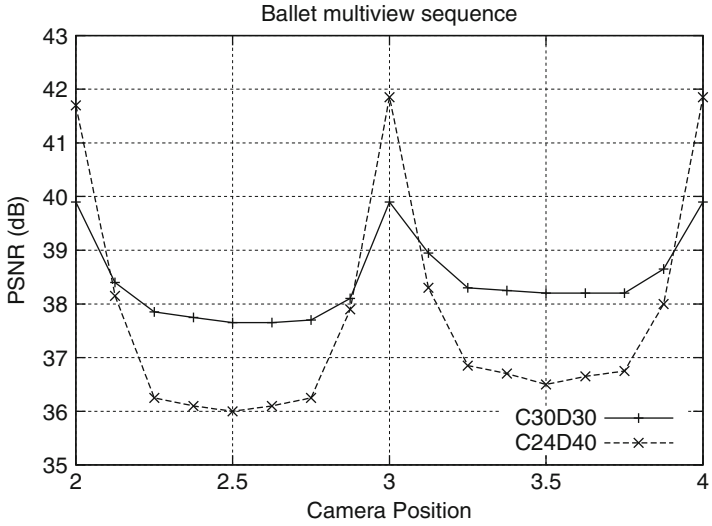


Fig. 2.22 PSNR across the viewing range of cameras 2, 3, and 4 for two different bit rate distributions between video and depth for the *Ballet* test set [26]

where video is decoded independent from depth, the decoder maintains compatibility with stereo decoders that do not support decoding of depth component. Otherwise, if view synthesis prediction is utilized, decoding of depth is required prior to decode the video [25]. Such tools have the potential to provide interesting compression gains at the cost of reducing compatibility.

An important issue in the design of joint video and depth coding is the quality optimization of synthesized views. Instead of evaluating the decoding quality in comparison with an uncoded reference, the MVD format, besides good video and depth quality, also requires good quality for the intermediate synthesized views. As often the original intermediate view is not present, comprehensive subjective evaluation is required. Such evaluation takes into account new types of errors, like pixel shifts, regions appearing with wrong depths, or outworn object boundaries at depth edges. In the experimental results presented in [26], PSNR measurement is used to compare the synthesized views with uncoded reference views. The *Ballet* test sequence is encoded using the overall bit rate of 1,200 kbps (video + depth), with two different pair of quantization parameters (QP) for both video and depth. One sequence is encoded with QP=30 for video and QP=30 for depth (C30D30). The other sequence is encoded with QP=24 for video and QP=40 for depth (C24D40), which means that the video bit rate is increased at the expense of the depth bit rate, thus increasing the video quality and reducing the depth map quality.

As can be seen in Fig. 2.22, at the original camera positions 2.0, 3.0, and 4.0, the configuration that uses a lower QP (24) achieves better reconstruction results than using higher QP (30). However, for the intermediate positions, "C24D40" performs worse than "C30D30" because it uses a higher QP (40) to encode the depth maps, thus degrading the depth signal quality and leading to displacement errors when using such depth data for view synthesis. Note that by using the depth information,

the view synthesis scheme warps the original views to an intermediate position and applies a view-dependent weighting to perform the view interpolation. As illustrated in Fig. 2.22, the furthest distance from any original view (2.5 and 3.5) presents the lower-quality values. Hence, these results show how important is to preserve the depth maps quality for the synthesis process, mainly at middle positions, far away from the original views. Besides these results, all dependencies between video and depth information are currently under evaluation in terms of their compression and rendering capabilities as well as the repercussions in the compatibility and complexity of future coding algorithms.

2.4 Asymmetric 3D Video Coding

Asymmetric 3D video coding relies on the binocular suppression theory of the human visual system (HVS), which states that a given stereoscopic content with different quality between the two views can be perceived with the same quality as that of the higher quality view [27]. One of the reasons for this consists in the HVS response, which does not notice as relevant the absence of high-frequency information in one of the views. Such characteristic of the HVS in regard to stereoscopic viewing can be used to achieve extra coding gains in comparison with classic single-view encoders where this is.

To benefit from the suppression theory, 3D video encoders may coarsely encode one component of the stereo video signal, in order to obtain increased coding efficiency gains without harming the quality experienced by users. Several methods can be used to achieve unbalanced quality between the two views. For instance, either the spatial or temporal resolution of one view might be reduced through preprocessing. The texture quality of one view can also be reduced by coarse quantization, resulting in asymmetric video encoding. Another method that can be used to reduce the amount of compressed data with little or no impact in subjective quality consists in dropping the chroma information of one view. Since chrominance degradation is less likely to be perceived, the chroma components of one stereo view can be dropped with no influence in the subjective depth perception of viewers [28]. Moreover if the lack of chroma information in one view is compensated through disparity or another reconstruction mechanism, then the color information in both views remains close to the original. This is because the image fusion process in the HVS superposes both images, resulting in a single chromatic content rather than two slightly different ones from each view.

2.4.1 Mixed Resolution for Asymmetric Coding

Asymmetric spatial resolution can be used in stereoscopic video coding by mixing different view resolutions [10]. Using mixed spatial and temporal resolution to a certain extent, the high-frequency information removed from one view by low-pass

Fig. 2.23 Asymmetric views obtained with low-pass filtering of one view [31]

spatial filtering is not detected by the HVS. In [10], spatial downsampling was implemented with filtering at 1/2 and 1/4 of the image resolution and temporal filtering was also done in two different ways, i.e., either by averaging the pixels from adjacent fields or dropping and repeating each other frame. The conclusions were drawn upon subjective testing carried out according to the ITU-R Recommendation 500 [29]. Other studies have also shown that mixed-resolution stereo video can achieve an overall perceptual quality and sharpness close to that of the higher quality view [30]. Therefore this is a possible method to reduce the coding rate of 3D video, providing that view asymmetry lies within an acceptable range. Figure 2.23 shows an example of asymmetric views obtained from low-pass filtering of one of them [31].

Mixed-resolution coding is also a valid option to reduce the bandwidth required by mobile 3D multimedia applications. In [31], the results of subjective tests found that for an overall bit rate of 400 kbps, the lower quality view can be encoded at 30–45 % of the total bit rate allocated to the stereo pair.

Although spatial resolution asymmetry is not supported in standard multiview encoders, this is still possible to implement using the scaling properties of scalable video coding (SVC). A possible coding strategy is to use non-scalable H.264/AVC for the base view while the auxiliary view is encoded with a modified SVC encoder [32]. Objective quality results show improved coding efficiency compared to simulcast (−52.05 Bjontegaard Delta (BD) bit rate or 4.66 BD-PSNR) or using inter-view predictions (−12.68 BD bit rate or 0.85 BD-PSNR). However since no subjective data is provided in [32], the actual perceptual performance of this method is not evaluated. To mitigate the additional computational complexity required for downsampling, spatial asymmetric multiview coding using lower complexity motion compensation can also be used [33].

Although mixed-resolution asymmetric methodologies proved to be an efficient, usable, and quality friendly method, similar studies on temporal-only asymmetric methods demonstrated low-quality levels. Some possible options to achieve temporal scaling of a stereoscopic video, such as dropping and repeating frames in one of the views or averaging frames to produce intermediate ones, yield unacceptable results [30, 34].

Fig. 2.24 Result of asymmetric quality coding

2.4.2 Asymmetric Quality

Using asymmetric quality in stereoscopic video provides an ease method to fine-tune the quality of rate-constrained video [30, 35]. Figure 2.24 shows an example of a stereo pair obtained from asymmetric coding where blocking artifacts can be seen in the right view but not so much in the left one. This type of SNR asymmetric coding is one of the most suitable to be used in current standard MVC encoders because it does not require coping with different resolutions and downsampling/upsampling operations.

Those views not used as reference for others can be encoded with higher QP or allocated a lower bit rate than other views. This results in asymmetric quality among primary (reference) and secondary (non-reference) views. It is also possible to interleave lower quality frames with high-quality ones in different views, e.g., odd frames in one view and even frames in another view with lower quality than the others [36]. This method has the advantage of eliminating viewer's asymmetric visual acuity (i.e., the influence of a dominant eye) and it is preferable to drop and repeat frames in the sequence.

A complete framework consisting of an MVD encoder and a bit allocation mechanism with chrominance reconstruction can be found in [37]. This approach consists in asymmetric coding of the MVD-based video using the JMVM codec with reconstruction at the decoder of those chrominance components that are discarded at the encoder in the lower quality views. Considering the total bit rate, this approach significantly improves the coding efficiency while maintaining the overall quality experienced by end users.

Another possible approach to encode asymmetric 3D video is to encode one view with a standard codec H.264/AVC and the other one with its scalable extension SVC [38]. Such scheme allows exploitation of a wide range of asymmetry with only one encoding channel. While the quality of one view is fixed, the one coded with SVC has high-quality scalability range and it can be easily extracted with either lower or higher quality than the H.264/AVC view.

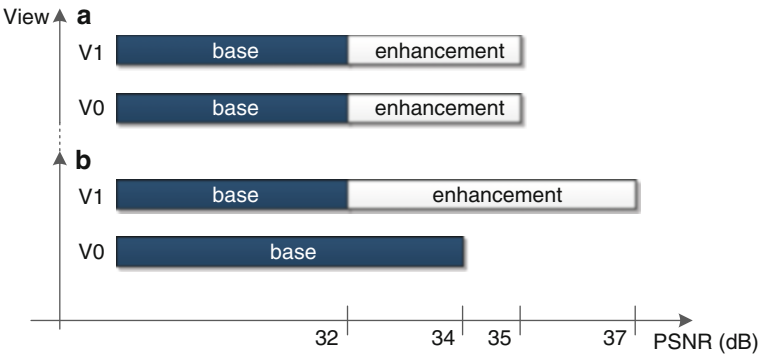


Fig. 2.25 Thresholds for asymmetric coding (**a**) both views coded with SVC and (**b**) only one view coded with SVC [38]

2.4.2.1 Perceptual Quality Thresholds

A relevant issue in asymmetric coding is to find perceptual thresholds beyond which quality degradation is noticeable by viewers. Relevant thresholds were experimentally found in [39] through subjective testing, where users started evaluating high-quality symmetric coding (i.e., PSNR = 40 dB) in both views and reducing the quality of the auxiliary view down to 25 dB. It was found that such threshold slightly varies according to the display and lies around 31 dB for a parallax barrier display and about 33 dB for full-resolution polarized projection displays. Thus an average value of 32 dB can be defined for the lower quality view when the other is very high quality. When comparing SNR scaling with spatial scaling, the former should be used at high bit rates (above the previously stated threshold) as it results in better perceived quality. When low bit rates are used (e.g., below 28 dB PSNR on the auxiliary view) spatial scaling tends to perform better than SNR scaling. When operating in some quality range between these two thresholds, symmetric coding is preferable over other options. Note that display technology also influences the subjective quality obtained by symmetric/asymmetric coding [40].

An indication of objective thresholds that can be used in asymmetric encoding of stereoscopic video is shown in Fig. 2.25, obtained from the results presented in [38]. In the first case two SNR scalable streams are produced and then the enhancement layer on one of them might be cut down to the threshold of 32 dB. In the other option, only one view is encoded with SNR scalability providing higher quality variation range, as mentioned before.

2.4.2.2 The Effect of Dominant Eye

Although asymmetric coding algorithms can be used to reach greater coding efficiency at either small or no cost to the viewer's perceived quality, when a lower quality view corresponds to the viewer's dominant eye, some users can

perceive the global quality lower than others [41]. Therefore, such ocular dominance effect may lead to an overall perceived quality not equal to that of the higher quality view as commonly expected.

To reduce the impact of the ocular dominance effect, it is possible to cross-switch the low- and high-quality views along the time [42]. However the interleaving of the views' quality along the time can also result in a noticeable effect, similar to flickering. To mitigate this effect, an adaptive algorithm should be implemented in asymmetric encoders such that views' quality cross-switching occur at scene cuts, where a perceptual masking effect of the HVS occurs. The GOP can also be used as a reference for cross-switching, as the frames inside a GOP have high correlation [43]. Even so the GOP size should not be small to avoid flickering effects. A possible alternative is to use unbalanced coding in horizontal slices with smoothing on the slice edges, in both views [44]. In this case, the high- and low-quality slices in each view should be located in complementary spatial positions in order to mitigate the effect of eye dominance.

2.4.2.3 Coding Regions with Different Perceptual Relevance

A finer approach to achieve spatial asymmetric coding is based on identification and encoding different regions of the stereo images according to their perceptual relevance. Relevant regions might be identified through a combined approach of depth variation thresholds and specific texture detection in order to differentiate the background and foreground of a 3D scene. Bit rate savings of 28 % can be achieved by using a method where the scene background is given less relevance [45].

Several approaches can be used to define regions of perceptual relevance. For instance, a threshold limit can be used to split either the depth values or disparity between stereo pairs in order to generate two different image regions. Stereo image regions with different perceptual relevance can be found by using a binary mask based on disparity thresholds between each stereo image [46]. An example of such masks is shown in Fig. 2.26. This method, when used as an extension of spatial quality asymmetric coding, is capable of yielding bit rate savings of about 20 % for the auxiliary view without noticeable subjective quality degradation.

2.5 Conclusion

This chapter addressed the most relevant representation formats and coding methods for 3D video services and applications. A thorough description was provided from the simplest frame compatible formats to the most complex ones, such as the MVD. The main characteristics of standard coding methods used for the various 3D video formats were also described, highlighting the additional tools specifically introduced to better cope with 3D visual content. Then asymmetric

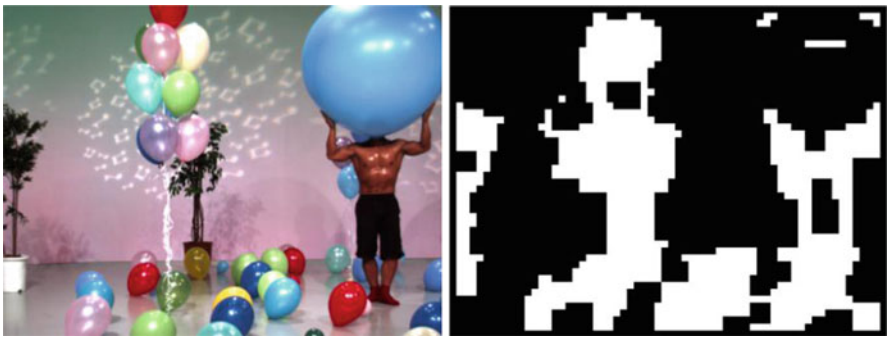


Fig. 2.26 Frame from sequence Ballons (*left*) and the corresponding mask identifying regions with different perceptual relevance for asymmetric coding (*right*)

coding techniques, relying on the particular characteristics of the HVS supported by the suppression theory, were also described. Several possible dimensions of exploiting asymmetric coding were highlighted, based on recent research findings. Overall this chapter provides useful information for anyone interested in this fast-evolving technology field of 3D video representation and coding.

References

1. Speranza F, Renaud R, Vincent A, Tam WJ (2012) Perceived picture quality of frame-compatible 3DTV video formats. In: IEEE international conference on multimedia and expo (ICME). pp. 640–645, Melbourne, July 2012
2. ITU-T and ISO/IEC JTC 1 (2010) Advanced video coding for generic audiovisual services. In: ITU-T Recommendation H.264 and ISO/IEC 14496–10 (MPEG-4 AVC)
3. Text of ISO/IEC MPEG2011/N12543 (2012) Additional profiles and SEI messages. San Jose, Feb 2012
4. Ballocca G, D'Amato P, Grangetto M, Lucenteforte M (2011) Tile format: a novel frame compatible approach for 3D video broadcasting. In: IEEE international conference on multimedia and expo (ICME), pp. 1–4, Barcelone, July 2011
5. Muller K, Merkle P, Wiegand T (2011) 3-D video representation using depth maps. Proc IEEE 99(4):643–656
6. Lee E-K, Ho Y-S (2011) Generation of high-quality depth maps using hybrid camera system for 3-D video. J Vis Com Image Rep 22(1):73–84, ISSN 1047–3203
7. Fehn C (2004) Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3DTV. In: Proceedings of SPIE stereoscopic displays and virtual reality systems XI, pp. 93–104, San Jose, Jan 2004
8. Smolic A, Mueller K, Merkle P, Kauff P, Wiegand T (2009) An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution. In: Picture coding symposium, pp. 1–4, Chicago, May 2009
9. ITU Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) ISO/IEC JTC1/SC29/WG11 (2013) High efficiency video coding. ITU-T Recommendation H.265 and ISO/IEC 23008–2 (HEVC)

10. Stelmach L, Tam WJ, Meegan D, Vincent A (2000) Stereo image quality: effects of mixed spatio-temporal resolution. IEEE Trans Circ Sys Video Technol 10(2):188–193
11. Tam WJ, Stelmach LB, Speranza F, Renaud R (2002) Cross-switching in asymmetrical coding for stereoscopic video. In: Stereoscopic displays and virtual reality systems IX, vol 4660, pp. 95–104, Burlingame, CAL, Jan 2002
12. Report ITU-R BT.2017, Stereoscopic Television MPEG-2 Multi-view Profile, 1998
13. ITU-T and ISO/IEC JTC 1 (1996) Final Draft Amendment 3. Amendment 3 to ITU-T Recommendation H.262 and ISO/IEC 13818–2 (MPEG-2 Video), MPEG document N1366, Sept 1996
14. Tian D, Pandit P, Yin P, Gomila C (2007) Study of MVC coding tools. Joint Video Team, Doc. JVT-Y044, Shenzhen, Oct 2007
15. Chen T, Kashiwagi Y (2010) Subjective picture quality evaluation of MVC stereo high profile for full-resolution stereoscopic high-definition 3D video applications. In: Proceedings of IASTED conference on signal and image processing, Maui, HI, Aug 2010
16. Su Y, Vetro A, Smolic A (2006) Common test conditions for multiview video coding. Joint Video Team, Doc. JVT-U211, Hangzhou, Oct 2006
17. Bjontegaard G (2001) Calculation of average PSNR differences between RD-curves. ITU-T SG16/Q.6, Doc. VCEG-M033, Austin, TX, April 2001
18. Jeon Y-J, Lim J, Jeon B-M (2009) Report of MVC performance under stereo condition. ITU-T & ISO/IEC JTC1/SC29/WG11 Docs. JVT-AE016, London
19. Chen T, Kashiwagi Y, Lim CS, Nishi T (2009) Coding performance of Stereo High Profile for movie sequences. ITU-T & ISO/IEC JTC1/SC29/WG11 Doc. JVT-AE022, London
20. Sullivan GJ, Ohm J-R, Han W-J, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. In: IEEE trans. circuits and systems for video technology, Dec 2012
21. Bossen F, Bross B, Sühring K, Flynn D (2012) HEVC complexity and implementation analysis. In: IEEE trans. circuits and systems for video technology, vol 22, no. 12, Dec 2012
22. Domański M, Grajek T, Karwowski D, Klimaszewski K, Konieczny J, Kurc M, Łuczak A, Ratajczak R, Siast J, Stankiewicz O, Stankowski J, Wegner K (2012) Coding of multiple video +depth using HEVC technology and reduced representations of side views and depth maps. In: Picture coding symposium, Krakow
23. ISO/IEC 23002–3:2007 (2007) Information technology—MPEG video technologies—Part 3: representation of auxiliary video and supplemental information
24. Vetro A, Tian D (2012) Analysis of 3D and multiview extensions of the emerging HEVC standard. In: SPIE conference on applications of digital image processing XXXV, pp. 8499–8433, San Diego, Aug 2012
25. Yea S, Vetro A (2009) View synthesis prediction for multiview video coding. In: Signal processing: image communication, vol 24, no 1+2, pp. 89–100, April 2009
26. Vetro A, Tourapis A, Karsten Müller, Chen T, 3D-TV content storage and transmission. IEEE transactions on broadcasting, Special issue on 3D-TV horizon: contents, systems and visual perception, vol 57, no 2, pp. 384–394, June 2011
27. Julesz B (1971) Foundations of cyclopean perception. University Chicago Press
28. Shao F, Jiang G, Wang X, Yu M, Chen K (2010) Stereoscopic video coding with asymmetric luminance and chrominance qualities. IEEE Trans Consum Electron 56:2460–2468
29. Recommendation BT.500-12 (2009) Methodology for the subjective assessment of the quality of television pictures, Sept 2009
30. Tam WJ (2007) Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV. JVT-W094, San Jose, CA
31. Brust H, Smolic A, Mueller K, Tech G, Wiegand T (2009) Mixed resolution coding of stereoscopic video for mobile devices. In: 3DTV conference: the true vision—capture, transmission and display of 3D video, pp. 1–4, Potsdam, May 2009
32. Quan J, Hannuksela M, Li H (2011) Asymmetric spatial scalability in stereoscopic video coding. In: 3DTV conference: the true vision—capture, transmission and display of 3D video (3DTV-CON), pp. 1–4, Antalya

33. Chen Y, Liu S, Wang Y-K, Hannuksela M, Li H, Gabbouj M (2008) Low-complexity asymmetric multiview video coding. In: IEEE international conference on multimedia and expo, pp. 773–776, Hannover, June 2008
34. Aksay A, Bilen C, Kurutepe E, Ozcelebi T, Akar GB, Civanlar MR, Tekalp AM (2006) Temporal and spatial scaling for stereoscopic video compression. In: EURASIP European signal processing conference, Florence, Sept 2006
35. Stelmach L, Tam W, Meegan D, Vincent A, Corriveau P (2000) Human perception of mismatched stereoscopic 3D inputs. In: International conference on image processing, vol 1, pp. 5–8, Vancouver, Sept 2000
36. Jain A, Bal C, Robinson A, MacLeod D, Nguyen T (2012) Temporal aspects of binocular suppression in 3D video. In: Sixth international workshop on video processing and quality metrics for consumer electronics, Scottsdale, Jan 2012
37. Shao F, Jiang G, Yu M, Chen K, Ho Y-S (2012) Asymmetric coding of multi-view video plus depth based 3-D video for view rendering. IEEE Trans Multimed 14:157–167
38. Saygili G, Gurler C, Tekalp A (2011) Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming. IEEE Trans Broadcast 57:593–601
39. Saygili G, Gu Andrler C, Tekalp A (2010) Quality assessment of asymmetric stereo video coding. In: 17th IEEE international conference on image processing (ICIP), pp. 4009–4012, Hong Kong, Sept 2010
40. Saygili G, Gurler C, Tekalp A (2009) 3D display dependent quality evaluation and rate allocation using scalable video coding. In: 16th IEEE international conference on image processing (ICIP), pp. 717–720, Cairo, Nov 2009
41. Reiss M, Reiss G (1997) Ocular dominance: some family data. Laterality 2:7–16
42. Tam WJ, Stelmach LB, Subramaniam S (2001) Stereoscopic video: asymmetrical coding with temporal interleaving. In: Proceedings of SPIE 4297, Stereoscopic displays and virtual reality systems VIII, pp. 299–306, San Jose, Jan 2001
43. Liu S, Liu F, Fan J, Xia H (2009) Asymmetric stereoscopic video encoding algorithm based on subjective visual characteristic. In: International conference on wireless communications signal processing, pp. 1–5, Nanjing, Nov 2009
44. Azimi M, Valizadeh S, Li X, Coria L, Nasiopoulos P (2012) Subjective study on asymmetric stereoscopic video with low-pass filtered slices. In: International conference on computing, networking and communications, pp. 719–723, Maui, Hawaii, Jan 2012
45. Karlsson L, Sjostrom M (2008) Region-of-interest 3D video coding based on depth images. In: 3DTV conference: the true vision—capture, transmission and display of 3D video, pp. 141–144, Istanbul, May 2008
46. Pinto L, Assuncao P (2012) Asymmetric 3D video coding using regions of perceptual relevance. In: International conference on 3D imaging (IC3D), Liège, Dec 2012

Chapter 3
Merging the Real and the Synthetic in Augmented 3D Worlds: A Brief Survey of Applications and Challenges

Athanasios M. Demiris

Abstract Augmented reality finds its way into everyday applications providing appealing alternatives to current interaction paradigms as well as efficient and comprehensive information visualization. The work presented in this brief survey was carried out in the earlier days of augmented reality research in three different application areas that show promising application potential for augmented reality and are still not completely addressed and solved. The survey addresses not only different application areas but through them different technological challenges. The application domains addressed are the valorization of cultural heritage through augmented reality, transformation of the simple TV viewing experience into interactive sports television, and visual support of interior design. In each of these application areas, specific technological problems of augmented reality are addressed and the initial approaches to their solution are presented, as well as a projection to current developments and potential adaptations.

3.1 Introduction

Augmented reality is not new. It has been conceived very long ago, while research in the realm has started in the 1960s, finding a wider basis in the 1990s. For the origins of augmented reality, many articles refer to the "Wonderful wizard of Oz" of 1901 as a very early source of many references to reality augmentation. Only 25 years later a magazine of the time, called "Radio News," publishing various very progressive title drawings introduces a kind of augmented reality application showing a semitransparent display depicting an opera singer streamed right into the living room of the time [1]. Most researchers agree that the first scientific mention of augmented reality technology was carried out by Sutherland, who

A.M. Demiris (✉)
Micro2gen Ltd., Technology Park NCSR Demokritos, Agia Paraskevi, Greece
e-mail: dema@micro2gen.com

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_3, © Springer Science+Business Media New York 2014

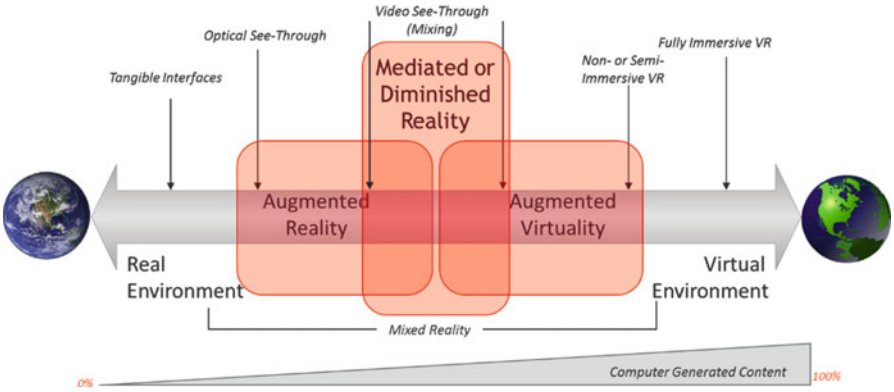


Fig. 3.1 A depiction of the reality-virtuality continuum based on the model introduced by Milgram and Kishino [5]

covers both virtual and augmented reality in his papers and ambitiously states that someday computers may be able to control the existence of matter [2]. His research work led to the "sword of Damocles," the very first head-mounted display aiming at an immersive experience for the viewer. In the 1980s the display of information projected right into the view field of fighter pilots combined with spatial sound was introduced. But it was only in the early 1990s that the term augmented reality was used for the first time by Caudell and Mizell of Boeing, who developed an AR system to support the aircraft technicians [3]. In the same year (1992) Steven Feiner, Blair MacIntyre, and Doree Seligmann presented KARMA, in the first extended paper on an AR prototype, in the framework of the Graphics Interface Conference and 1 year later in [4]. In 1994 Milgram and Kishino presented the reality-virtuality continuum in [5], which is depicted in the Fig. 3.1. The continuum model introduces various levels of integration between real and synthetic content and positions the terminology on this axis. A taxonomy, further refining parts of the reality-virtuality continuum that was introduced specifically for augmented reality and can be used to classify augmented reality systems, can be found in [6].

Augmented reality is possible only through the collaboration of different engineering disciplines and provides increased realism mainly through complicated setups that require an optimal interplay of processing and display hardware, sensors, and software processing. In order for an optimal result to be presented to the end user, collaboration with other (nonengineering) disciplines is required and varies based on the application under consideration. For example, in order to create successful augmented reality applications for museums and archaeological sites, it is necessary to at least involve museologists, designers, and eventually directors to best "shape up" the necessary storytelling. Most augmented reality applications in commercial use to date are applying mainly an augmented understanding of reality,

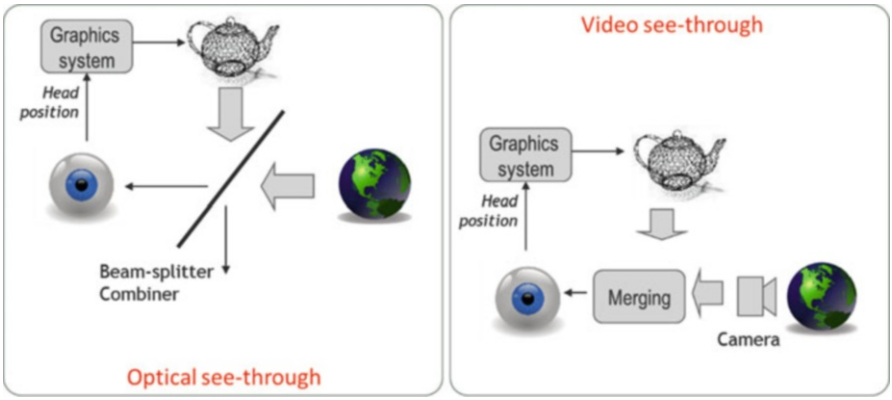


Fig. 3.2 Depiction of the basic working concept of optical (*left*) vs. video see-through displays (*right*). The eye depicts the user, the globe the real content, and the teapot grid the synthetic content

but in this paper we will focus on the perceptual and partly behavioral association of the real and the virtual, as well virtualized reality, following the taxonomy in [6].

The present paper focuses mainly on software aspects, although in many cases, the sensors available are also mentioned, since they influence the accuracy of measurements. But the display used to produce the final impression of augmentation plays a central role in every augmented reality application. The display hardware at hand is an important factor determining not only the user experience but also the type of software processing itself. As shown in Fig. 3.2, the two main types of portable displays are optical see-through displays and video see-through displays (the list of technologies for fixed displays that can be used as "ambient" installations in an augmented reality experience is very long and not dealt with in this context). In head-mounted or portable optical see-through, a real image of the world is presented to the end user using appropriate mirrors and optics. The major disadvantage is that the synthetic objects are usually projected on the basis of small displays integrated in the setup and usually a rectangular area corresponding to these displays becomes visible leading to a difficult to control and partially unnatural outcome. A potential solution would be to project the display in an area larger than the view field and to take that into account when calculating the synthetic content, but no such implementation was available until 2005. In video see-through the image projected does not come directly from the real world through optics, but through a camera. As such each frame becomes available in the memory of the processing unit and the synthetic objects are overlaid there prior to projection.

In the following sections, three major technological challenges of augmented reality applications are presented. Three case studies are selected not only for the reason that they provide very attractive application scenarios for augmented reality but also because they highlight important and only partially solved problems for the wider spreading of augmented reality. A major motivation behind the selection of

the survey cases is also that today, approximately 10 years after the implementation of the projects described, even commodity hardware, has evolved dramatically providing not only significantly increased processing power but also access to numerous sensors delivering very accurate measurements of position, orientation, velocity, light, etc., and the same problems may be tackled in a much more efficient way.

In the following sections, each application area is presented in its entirety but with special focus on a specific technological problem that is representative for the given case. Hence, in the case of cultural heritage, markerless tracking plays a central role, since ancient sites usually need to remain unaltered; in the case of interactive sports television, the near-real-time transmission of human motion in a sports event (also without any body markers) onto a synthetic avatar is addressed, and finally the creation of photorealistic, merged views of the real and the synthetic plays a crucial role in interior design and decoration. This paper focuses on the presentation of the augmented reality challenges addressed by three collaborative European research projects about one decade ago and does not cover a complete survey of augmented reality applications. For a more complete survey, there are numerous online and printed reviews, such as the one in [7].

3.2 Augmented Reality in Cultural Heritage: The Problem of Markerless Tracking

One of the most obvious and widely spread applications of augmented reality is in the valorization of cultural heritage and particularly archaeological and historical sites. The use of augmented reality is expected to boost the "readability" of these sites by projecting onto the actual environment information that may help visitors better comprehend the historical value and immerse into the life of past civilizations. Different levels of augmentation may lead to different levels of readability, starting by simple projections of artifacts in their actual surroundings, to reconstructions of ancient temples and buildings all the way to the reconstruction of life in an ancient site.

Although many projects exist in the realm, the survey is going to focus on the work carried out in the framework of the European research project ARCHEOGUIDE [8], which aimed at the revival of the site of ancient Olympia in Greece, using mobile augmented reality guides. The idea was to create personalized tours through the site and have the system guide the visitor from place to place providing both visual and audible enhancements that would help understand the way the site was built and functioning in its prime years in ancient Greece. The augmented tour featured reconstruction of monuments in situ, completing the ruins on-site, but also revival of events, such as races in the stadium or religious offerings in the temple area. Important artifacts, such as the never found, oversized statue of Zeus in the position of the ruins of the corresponding temple,

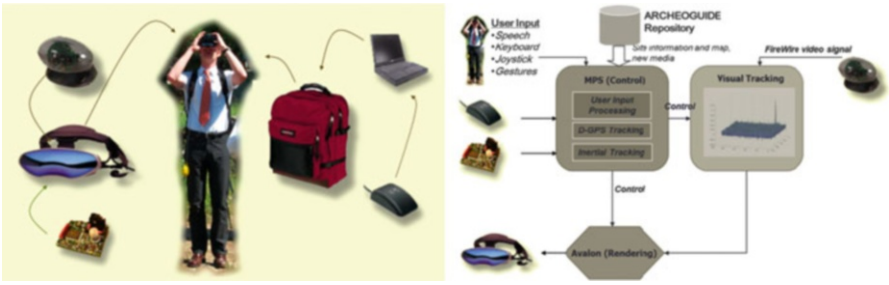


Fig. 3.3 The components and peripherals in the AR system and their connectivity

could be displayed to the visitor according to the archaeologists' views, in order to help the visitor better understand the importance of the site.

The usage scenario supports the nomadic behavior of the end users and foresees pickup of the augmented reality guide by the visitors at the entrance of the site. The guide was available in three variants: (1) a complete portable augmented reality system with a processing unit, video see-through binoculars, a headset, and a variety of sensors, all positioned in a compact backpack (except for the binoculars, which were handheld, and the headset, which was placed on top of one ear); (2) a smart book, which featured all necessary peripherals; and (3) a pocket guide, based on a PDA (personal digital assistant, the predecessor of the smart phones). Firstly the visitor is asked to provide some information in order to personalize the tour experience and an appropriate tour is created for the specific visitor. Through audio guidance the visitor is guided to different areas within the site. Every time multimedia content was available, the visitor was prompted to look through the binoculars. The presentation was in most cases augmented into the scenery.

A very important prerequisite was that there should be no alteration in the site of any nature or extend. Even very small markers to be used as fiducial points by the system were not allowed and were considered as intervention. This led clearly to the need for a markerless tracking approach.

The equipment used was a laptop with a mobile graphics card in the backpack, equipped with an assisted GPS sensor, a gyroscope, a camera, and a joystick for the user interactions. The binoculars featured a set of buttons that could be used alternatively to the joystick. Most sensors were built into or attached to the binoculars, since they serve in the determination of the optical view field of the visitor. The laptop featured a wireless LAN connection in order to be able to receive data depending on the location and type of tour guidance taking place (Fig. 3.3).

3.2.1 Use of Reference Images

In the core of the system lies a multimedia database that contains all content in a geo-tagged manner. For each significant position in the site, there is a series of different multimedia files that can be used to either display or playback contextual

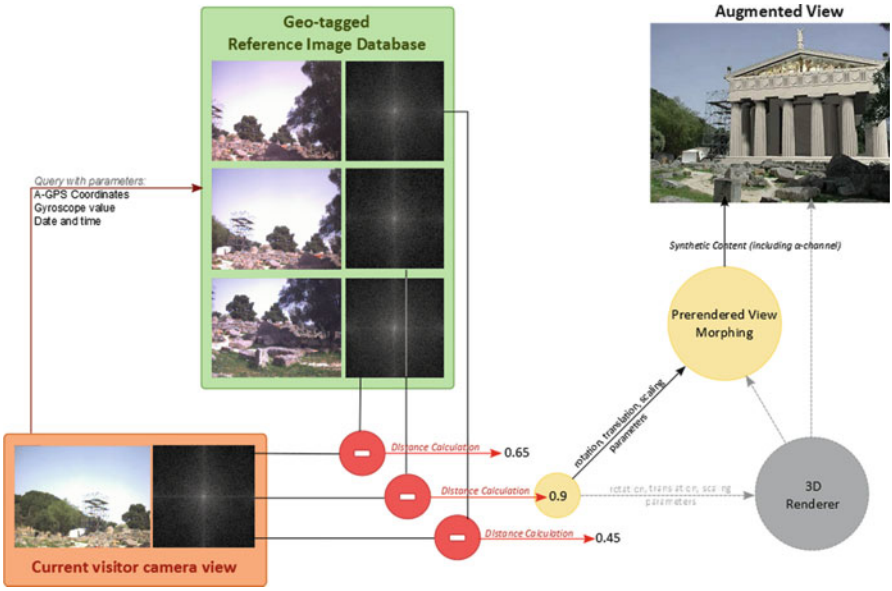


Fig. 3.4 An overview of the reference image usage for the calculation of the projection parameters leading to the final augmented view; two possibilities for obtaining the final rendering are presented (the most time- and power-consuming one depicted in *gray* with *dotted lines*)

information or augment the scene. Among these files are also reference images that are used to support the tracking of the scenery. Since no markers can be used, the solution provided was to use different images of the environment and calculate a visible distance between them and the footage from the camera.

The images exist for different times of the day and different seasons, since there are significant changes in the scenery caused by very dark shadows cast by the strong sunlight in the morning and smaller and softer shadow areas in the afternoon. The environment in ancient Olympia is exhibiting very strong variations between different seasons, especially spring and summer (green vs. dried). This has an impact on the distance calculations between the reference image and the images coming from the camera mounted on the binoculars.

In order to calculate accurate distances between the images, metrics are necessary, which would be rotation, translation, and scaling invariant, since the free movement of the visitors' hands holding the binoculars is expected to create variations in all axes, in addition to the different scaling due to the distance of the visitor to the object. Stricker has applied in [9] a Fourier-Mellin transform to recover rotation, translation, and scaling and perform a registration between two images. In the methodology applied, first the fast Fourier transform (FFT) of the two images to be compared is calculated (for reference images in the database, these are precalculated). Subsequently a log-polar transformation of the spectrum takes place. The phase-correlation method is then applied and the rotation angle as well as the scale factor is recovered. At the end, and after image rectification, the translation is calculated, once again by phase correlation (Fig. 3.4).

The rotation, translation, and scaling parameters can be used to project the synthetic object onto the current frame from the camera footage. This can be done in two ways: a more time-consuming but accurate method is to feed a renderer with the parameters and create a new view for the current position of the visitor and then project it onto the image and have the augmented view displayed on the video see-through setup. The second option, which requires significantly reduced resources when compared to the first one, foresees the morphing of an existing pre-rendered view onto the current frame. This usually works very well with very good perceived quality results, since a large number of different reference images used for different fields of view and in addition the areas where the display is triggered (by creating and audible signal for the visitor that there is visual information available in that particular spot) are relatively small, so that the variance in scaling, rotation, and translation is marginal.

3.2.2 Projection to Current Developments

Markerless tracking is still a challenging topic in machine vision and augmented reality. There are numerous applications that would benefit from a stable markerless tracking solution, and although many domain-specific solutions exist, there are no generalized solutions available.

The implementation of markerless augmented reality for significant (or even less significant) places based on reference images is becoming feasible thanks to the digital maps that are widely available (e.g., the "street-view" feature of Google Maps). Currently augmented reality applications similar to the one described are feasible for multiple sites and cases beyond cultural heritage valorization. Most portable devices are equipped with all sensors necessary to make a good position and view-field orientation of the user, along with a camera capture, which allows for the markerless tracking described above.

3.3 Augmented Reality in Interactive Sports Television: Projecting Human Motion onto Avatars

The second application presented is related to augmented, interactive television for sports broadcasting. Sports broadcasting is the ideal case for the application of interactive scenarios, where a viewer will most likely want to retrieve additional information about an event in different formats, while it is taking place. A merging of interactive computer games and live sports broadcasting leads to a much better comprehension of sports events and various scenarios such as three-dimensional reconstructions of attempts of athletes, where navigating freely in the scene by the viewer becomes feasible. This merging of real and synthetic content requires very precise extraction of various parameters from the real motion in real time or near real time, such as human body motion, while others may be prepared offline beforehand (e.g., the avatar of the athlete).

In the following sections, the requirements and techniques are presented, which were applied in the collaborative European research project PISTE [11], co-funded by the European Commission in its 6th Framework Programme. The creation of interactive sports television in real time or near real time (only seconds or minutes after an attempt, an interactive replay becomes available) requires adaptations in the acquisition of footage, as well as pre- and postprocessing of the scene and the attempt, in order to create an augmented view of the event. Due to the complexity of the overall effort, the project focused on track-and-field sports with just one contestant (primary focus was on high jump, pole vault, and long jump). This significantly simplifies the determination and reconstruction of human motion.

The enhancements offered to the viewers of a sports event are related to projecting additional information onto the actual footage, such as speed measurements, trajectory highlights, comparative display of multiple attempts, but also the ability to freely navigate within the scenery, while replaying an attempt of an athlete. In all cases the environment needs to be captured in a formal representation and after the necessary image segmentation, it needs to be combined with visualizations based on the motion of human athletes.

For each of the broadcasting cameras, a panorama image covering the field of view of this camera is computed. In the approach by Fraunhofer, IGD [10] images from multiple camera positions are incorporated for aligning the images. This approach leads to correct 3D calibration data in reference to the coordinate system of the reconstructed 3D scene. Correspondences between images from different camera positions have to be set manually, while the alignment of the video sequences from the same position into a panorama image is performed fully automatically. Several camera models are supported for the environment map: the spherical coordinate system is the classical coordinate system for panorama images, but requires costly mathematics for the mapping and cannot easily be used for full-view panoramas because of the singularities at the poles. Alternatively, the environment can be mapped onto the six faces of a cube, which leads to a simplification in the mappings. A third possibility is the use of the classic pinhole model of a perspective camera, which is used for original camera images in case of static cameras.

The calibration information associated with this so-called environment map makes it possible to compute the camera's viewing ray in 3D space for each pixel position in the image. Consequently, images showing the athlete from the same camera position just need to be 2D aligned with the environment map in order to calculate the 3D calibration parameters for this image. The advantage is the low processing cost, the independency of each calibration step from other images, and the ability for parallelization.

3.3.1 Offline Data Acquisition

The data acquisition takes place offline both before the event and during the event. The preparatory phase aims at the collection of all information necessary to reconstruct the event and the settings it takes place in. The three-dimensional

model of the environment is reconstructed from multiple photographs using photogrammetric techniques [12]. In the technique introduced by Fraunhofer IGD, the user is solely required to identify the corners of the objects to be reconstructed in the photographs and to select the faces that connect these corners and form a 3D model of the correct topology. With this input, the calibration of the cameras is performed so that their position, direction, and lens parameters are captured. Subsequently, the 3D positions of the corners are computed and the surface texture is extracted from the photographs automatically. By including example images from each TV camera in this process, the positions and lens parameters of the TV cameras are estimated as well, while additional images from the same camera position can be included fully automatically.

This approach is clearly aiming at a result with predominant synthetic content featuring real-world characteristics (such as the textures and the human motion, to be described in the following). In the category of offline preparatory activities also falls the collection of all accompanying information, such as schedule, names of athletes, nationality, and various demographics and statistics. The schedule and name of athletes can be used to pre-fetch avatars that are customized to match the bodies and look of the specific athletes. Ideally each athlete participating in the event would have their avatar prepared after a full-body scan in the beginning of the event.

3.3.2 Online Data Acquisition and Pose Estimation

The creation of interactive three-dimensional views of an attempt requires capturing the event from multiple cameras, at least two, simultaneously. These cameras need to be clock synchronized in order to produce frames (or fields in case of analog cameras) with identical timestamps. Subsequently each footage is processed complementary to the others. For each frame (or field), the additional frames (or fields) from the other cameras need to be processed in parallel.

The processing steps (vision pipeline) basically consist of multiple iterations of (1) silhouette determination through segmentation (in the case of PISTE through a modified/enhanced region-growing algorithm), (2) pose estimation (in this case through a statistical model), (3) pose correction (in this case deploying a three-dimensional anatomical atlas), (4) prediction of the pose for the next frame/field, and finally (5) projection of the predicted pose onto the different camera views, to be used as starting point for the subsequent segmentation.

It was decided early in the project to use 18 distinct significant anatomical points on the body of each athlete and determine their position in space in order to reconstruct the body motion. These 18 points, when appropriately connected, form a skeleton that is projected in three-dimensional space and back-projected to two dimensions for prediction of the subsequent pose (once again taking advantage of epipolar geometry techniques). These points were selected to coincide with a subset of the nomenclature selected and used in the MPEG-4 BDP standard (body

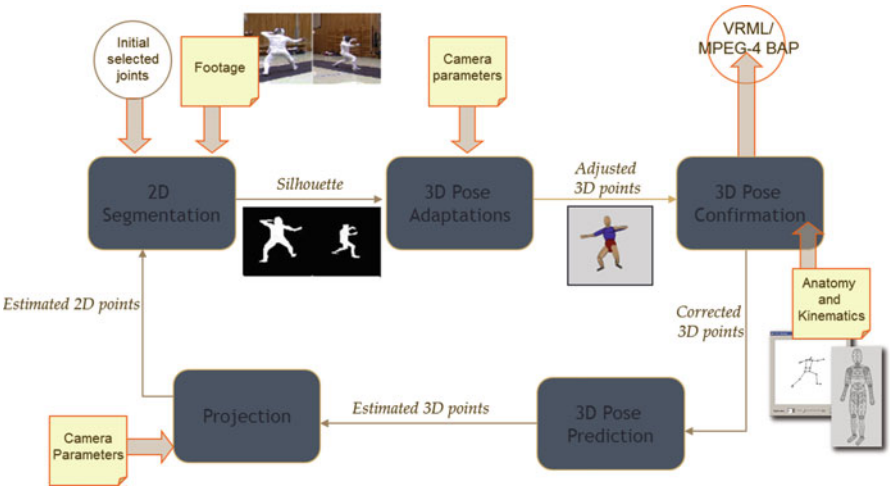


Fig. 3.5 The PISTE vision pipeline featuring segmentation algorithms by the University of Crete, pose estimation and adaptation by the Zentrum für Graphische Datenverarbeitung, Darmstadt, Germany, and 3D pose confirmation based on an anatomical model of the human by the University of Hannover, shown with sample images from a fencing event

description parameters). Correspondingly for the representation of the motion (of these selected body points), the MPEG-4 BAP (body animation parameters) standard was used. Obviously, not every point was visible from every camera, due to occlusion of body parts. In theory, it was sufficient to have all points detected by all cameras involved in the first frames and then cover for missing ones based on the anatomical model used. Nevertheless, in practice, it proved helpful to interrupt the automatic processing whenever the system could not accurately determine half the points and ask for user intervention. Hence, the solution proposed in the PISTE project was semiautomated. For a duration of approximately 10 s (or less) that is necessary to cover a high-jump attempt, several user interventions were necessary to correctly position body points and relaunch the process from that frame on (Fig. 3.5).

3.3.3 Point Distribution Model

For the description of the motion of the athletes, a statistical kinematic model was selected, namely, the point distribution model (PDM, as used by Heap et al. in conjunction with active shape models in [13]). PDM derives a statistical description of objects from training data sets and is particularly useful for nonrigid models. The statistical training necessary leads naturally to different kinematic models for each type of sports. The details of the implementation are described in [12]. Ideally a

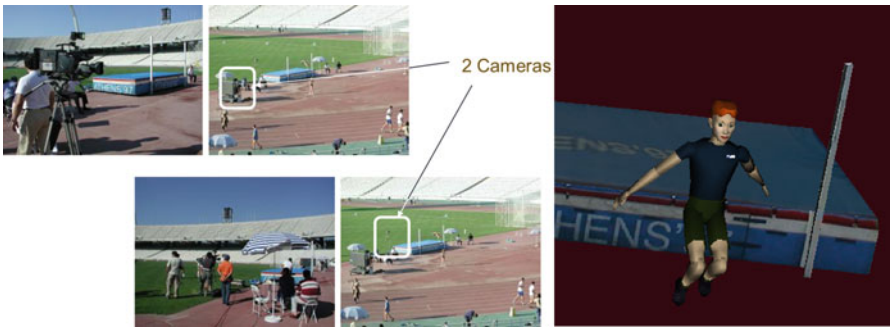


Fig. 3.6 On the *left* the setup for the high-jump event during the 2003 Panhellenic track-and-field games captured by two cameras of the Hellenic Broadcasting Corporation and on the *right* a randomly selected avatar performing the movement of a (female) athlete as calculated and rendered by the Zentrum für Graphische Datenverarbeitung and Fraunhofer IGD

large database could hold personalized versions of the PDM for each athlete based on past attempts in order to accommodate for any personal style in the movement. The system within PISTE was tested only using generalized models for each type of sports tested (Fig. 3.6). Usually the training set size would range between 20 and 40 different motion captures for each type of sports.

3.3.4 Prediction and Correction Using Anatomical Models

As mentioned before, in the vision pipeline of the project, after the estimation of the skeleton of the athlete for a particular frame from different synchronized cameras, and its projection in three-dimensional space, a validation and adaptation step follows, which is evaluating the result, using a generic 3D body model in order to identify unnatural movements or poses and possibly correct them or ask for user intervention. The generic anatomical model used consists of 15 simple volumetric primitives, representing significant body parts, which are fully described by the 18 important points selected for detection on the athletes' bodies. Approximate body proportions are taken from anthropometrical descriptions of human bodies, as described in [14].

The pose adaptation takes place in a hierarchical manner, starting with the torso, chest, and belly moving subsequently to other body limbs. The three-dimensional model is positioned according to the information coming from preceding steps related to the coordinates of the 18 aforementioned body points and then back-projected onto the two-dimensional view of each camera involved. The differences between the back-projected synthetic and the calculated actual silhouette are used to correct the position of individual points and then regenerate a three-dimensional pose, until the differences in all silhouettes drop below a selected threshold. In that

case the PDM is used to predict the next position of all points; a three-dimensional model is generated and adjusted to coordinates relative to the latest calculated pose. In the next step the silhouette for each camera is produced once again through back-projection. The result is used to produce the seed points for the multi-seeded region-growing algorithm used to segment the silhouette in each frame.

3.3.5 Other Enhancements and Projection to Current Developments

All processing presented thus far aims exclusively at the production of an interactive three-dimensional replay of an attempt in sports. As mentioned before, the motivation behind this type of enhancement and mixed-reality television is to provide viewers with a novel experience where they can navigate freely in a computer game-like manner assuming the role of the director and focus on the details and aspects they find most interesting. Mixed-reality television for sports can provide many more enhancements and solutions than three-dimensional replays, like visual overlays depicting additional information, such as the path that an athlete or an object has followed, various measurements like velocity and distance, or even comparison between different attempts in the same picture. Such features are very interesting not only for the interested viewer but also for the sports specialists, trainers, and the athletes themselves. The project took advantage of the three-dimensional reconstruction of the environment where the events took place and implemented some examples of such features. A very small subset has found its way to commercial television, e.g., during broadcasts of soccer games with the projection of various measurements and auxiliary lines onto the field. The interactive activation or deactivation of such enhancements through a smart set-top or TV set would lead to an initial form of content personalization in television.

3.4 Augmented Reality in Interior Design: Augmented Reality and Photorealism

A vital aspect in augmented reality, which will play a central role in the acceptance of augmented reality applications in everyday life, is photorealism. Many of the first attempts of augmented reality ended up in synthetic objects not really seamlessly integrated into the real scene, with very intense and unnatural colors projected on the real-world surrounding, not actually delivering the value expected from such applications. A very interesting application area for augmented reality is interior design. Augmented reality may be used by interior designers or architects to show their designs to their customers and together walk through empty spaces decorating

and populating them with furniture, discussing and experimenting with alternatives. For this type of application, photorealism is crucial.

The European collaborative IST Project ARIS, co-funded by the European Commission in its Fifth Framework Programme, dealt exactly with this topic between 2000 and 2004. The project investigated two application scenarios, the interactive version aiming at the visual selection of additional furniture for an already furnished room (e.g., e-commerce-site visitor who wants to see a piece of furniture in their environment prior to purchasing it and is willing to take a few snapshots using a webcam and have the augmentation take place on the servers of the provider) and a second real-time collaborative AR scenario, where groups of people jointly place objects in an empty space and modify selections (e.g., interior designers or architects with customers arranging furniture, lights, etc., in a new building). All results of the project related to photorealism were presented in [15] and the discussion in this section is based on this paper.

Photorealism in augmented reality requires a set of different light-matter interactions between the synthetic and the real. Apart from the correct occlusions, there should also be shadows cast from synthetic objects onto real objects and vice versa. In addition, the light in a real scene should be modeled and used by a renderer to produce the synthetic objects, as if they were in the scene, before integrating them in the actual settings through projection on a display. Finally, especially in interior decoration, there are cases where a synthetic light source may be casting light onto real objects. In such cases, when using video see-through displays, the brightness of the spots of real objects may be altered to reflect the additional illumination by the artificial object.

As described in all previous cases, the first step is the creation of a geometric model based on images of the scene. This can be done on the basis of one or more images of the scene. A technique for the calibration of the camera position and the generation of an approximate representation of the scene on just a single picture of the scene was used in the project for the desktop version (application scenario of adding furniture to an already furnished room) by INRIA, by extending the technique of vanishing points selection introduced in [16]. Assuming each pair of edges is not parallel, they can be intersected in image space and used to estimate the camera focal length, position, and orientation. Camera orientation is found by using the inverse of the camera calibration matrix to transform image-space vanishing points into direction vectors.

In order to accurately reconstruct the illumination conditions in the room (the so-called photometric scene reconstruction, as opposed to the geometric scene reconstruction), a technique was developed featuring high-dynamic-range images and a so-called spherical light probe positioned in the room prior to the "augmentation session." Based on the HDR images of the light probe, it is possible to project all light sources onto the surfaces of the 3D reconstruction of the scene, creating a so-called radiance mesh. This radiance mesh can then be used to calculate all shading and shadows cast onto and by synthetic objects (Fig. 3.7).

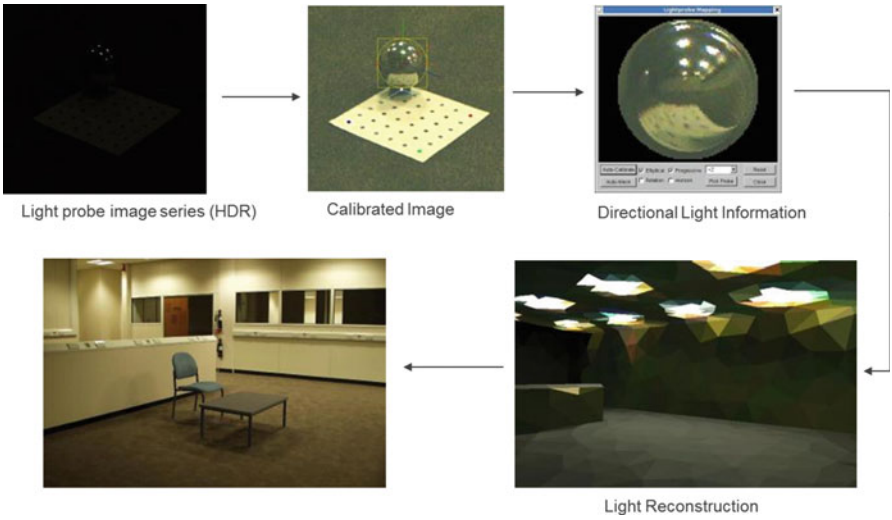


Fig. 3.7 The procedure to determine the light sources and create a radiance mesh for a given environment, in order to augment synthetic objects in a photorealistic manner (table and chair in the last image in the sequence) as introduced by the Simon Gibson of the University of Manchester in the ARIS project. The *dotted pattern* in the second image in the sequence is used to enhance the accuracy of the reconstruction by comparing the actual shadow cast by the light probe against the reconstructed one and calculating a potential shift

3.4.1 Current Developments

It is obvious that a working solution for all the above with minimal user interaction or preparation will benefit all augmented reality applications significantly increasing their perceptual quality for all users. Nevertheless, the state of the art has not yet exhibited many successful generic solutions that can be incorporated automatically into augmented reality solutions without the need for extensive user interaction. Hence, photorealism is possible, but requires some preprocessing and a procedure that may not be suitable for all types of users. It is thinkable that such an AR system could be provided as a service by an interior decoration company where trained personnel would take over all the necessary preparatory activities. Hence, photorealism is desirable for all augmented reality applications but at the time being may be deployable only in cases where it is inevitable due to increased preparatory overhead, such as interior design and architecture.

3.5 Conclusion

The work presented herein has taken place in collaborative projects in the realm of augmented reality a decade ago. Nevertheless, both the applications and the problems addressed by these projects are still attractive and not satisfactorily

solved, respectively. The three major problems addressed are markerless tracking, real-time human motion calculation and projection onto avatars, and photorealistic merging of the real and synthetic objects in augmented views. All these issues still play a central role in the acceptance of augmented reality applications and offer viable scenarios for the added value and widespread use of augmented reality. Current hardware is capable of significantly reducing the processing time necessary to fulfill real-time requirements, and developments in graphics hardware may lead to an increased perceptual acceptance of photorealistic AR. This contribution aims at triggering such developments and at providing a collection of references to techniques and approaches that have led to working solutions, where only additional engineering is required to produce commercial results using current hardware and development environments.

Acknowledgements The work presented here was carried out in the collaborative projects mentioned before that were co-funded by the European Commission in the 5th and 6th Framework Programme. The author participated in these projects as a member of the research and development team of Intracom S.A. either as project coordinator (PISTE) or coordinator of the company's team (ARIS) and/or member of the design and development team (ARCHEOGUIDE) and wishes to thank and mention the contributions of all colleagues that participated in these developments, especially the deputy director of the R&D department Nicolaos Ioannidis and the colleagues Maria Traka, Vassilios Vlahakis, Alexandra Makri, Ioannis Christou, and Ioannis Karigiannis. In addition the author would like to acknowledge the contributions of Fraunhofer Institute's Didier Stricker, Stefan Mueller, Georgios Sakas, Konrad Klein, Cornelius Malercyzk of the Zentrum für Graphische Datenverarbeitung, Prof. Krzysztof Walzack of the University of Poznan, Alan Chalmers and Patrick Ledda of the Bristol University, Simon Gibson of the University of Manchester, as well as colleagues from France Telecom (Alexandre Cotarmanac, Isabelle Marchal), Thomson Multimedia (Paul Kerbiriou), Jochen Wingbermühle and Torsten Wiebesiek of the University of Hannover, Emmanuel Reusens of Dartfish, AC2000, and all other partners in these projects. All scientists are mentioned using the names of the institutions they were with at the time of the project implementation, although many have later changed to different institutions or enterprises.

References

1. MagazineArt, Radio News Magazine. http://www.magazineart.org/main.php/v/technical/radionews/RadioNews1927-05.jpg.html
2. Sutherland IE (1965) The ultimate display. In: Proceedings of IFIP 1965, vol 2. p 506–508
3. Caudell TP, Mizell DW (1992) Augmented reality: an application of heads-up display technology to manual manufacturing processes. In: Proceedings of 25th Hawaii international conference on system sciences, vol 2. IEEE
4. Feiner S, Macintyre B, Seligmann D (1993) Knowledge-based augmented reality. Comm ACM 36(7):53–62
5. Milgram P, Kishino F (1994) A taxonomy of mixed reality visual displays. IEICE T Inf Syst E77-D(12):1321–1329
6. Hugues H, Fuchs P, Nannipieri O (2011) New augmented reality taxonomy: technologies and features of augmented environment. In: Furht B (ed) Handbook of augmented reality. Springer, Berlin

7. van Krevelen DWF, Poelman R (2010) A survey of augmented reality technologies, applications and limitations. Int J Virtual Reality 9(2):1–20
8. Vlahakis V, Karigiannis J, Tsotros M, Gounaris M, Almeida L, Stricker D, Gleue T, Christou IT, Carlucci R, Ioannidis N (2001) Archeoguide: first results of an augmented reality, mobile computing system in cultural heritage sites. In: Proceedings of the 2001 conference on virtual reality, archeology, and cultural heritage (VAST '01). ACM, New York, p 131–140
9. Stricker D (2001) Real-time and markerless vision-based tracking for outdoor augmented reality applications. In: IEEE and ACM international symposium on augmented reality. IEEE Press, Los Alamitos, California, p 189–190
10. McMillan L, Bishop G (1995) Plenoptic modeling: an image-based rendering system. In: Computer graphics proceedings, annual conference series (SIGGRAPH '95), ACM
11. Demiris AM, Traka M, Reusens E, Walczak K, Garcia C, Klein K, Malerczyk C, Ioannidis N (2001) Enhanced sports broadcasting by means of augmented reality in MPEG-4. In: International conference on augmented, virtual environments and three-dimensional imaging, Mykonos
12. Malerczyk C, Klein K, Wiebesiek T (2003) 3D reconstruction of sports events for digital tv. In: 11-th international conference in Central Europe on computer graphics, visualization and computer vision 2003. WSCG, Plzen
13. Heap T, Hogg D (1996) Towards 3D hand tracking using a deformable model. In: International conference on automatic face and gesture recognition, Killington, Vermont. IEEE Computer Society Press, Los Alamitos, Oct 1996
14. Tilley AR, Henry Dreyfuss Associates (2001) The measure of man and woman: human factors in design. Wiley, New York
15. Gibson S, Chalmers A, Simon G, Vigueras Gomez J-F, Berger M-O, Stricker D, Kresse W (2003) Photorealistic augmented reality. In: Second IEEE and ACM international symposium on mixed and augmented reality—ISMAR'03. IEEE, ACM, Tokyo, p 3
16. Cipolla R, Drummond T, Robertson DP (1999) Camera calibration from vanishing points in images of architectural scenes. In: Proceedings of the British machine vision conference, vol 2. Nottingham, p 382–391

Chapter 4
Multi-view Acquisition and Advanced Depth Map Processing Techniques

Nicolas Tizon, Gabriel Dosso, and Erhan Ekmekcioglu

Abstract This chapter provides a general framework for multi-view and 3D video content acquisition. The device architecture is described as well with the camera post-processing stage, taking into account of the multi-view aspects. In addition, a depth-map computing stage is described in order to provide a complete specification of the multi-view content acquisition and preparation module. For the depth map extraction, different levels of quality can be achieved depending on the processing time. For monitoring purposes, the priority is given to the fastness although the depth information supposed to be used for the rendering, at client side, requires more computing in order to optimize the quality of the reconstructed views. In addition, in the multi-view context, a refinement step aims at enhancing the quality of the depth maps and optimizing the video coding performances. Finally, experimental results are presented in order to validate the different approaches through quality measurements of depth-based view reconstruction.

4.1 Introduction

The delivery of 3D immersive media to individual users is becoming a key requirement when designing new advanced multimedia applications or services. Beyond the basic glasses-based stereoscopic rendering, the last researches on the topic are more oriented toward multi-view rendering [1, 2]. In this direction, one can distinguish between two main technologies:

N. Tizon (✉) • G. Dosso
VITEC Multimedia, 99 rue Pierre Sémard, 92324 Châtillon, France
e-mail: nicolas.tizon@vitecmm.com; gabriel.dosso@vitecmm.com

E. Ekmekcioglu
I-Lab Multimedia Communications Research Group, University of Surrey,
Guildford, GU2 7XH Surrey, UK
e-mail: erhan.ekmekcioglu@surrey.ac.uk

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_4, © Springer Science+Business Media New York 2014

- Auto-stereoscopic 3D rendering: the 3D display shows stereoscopic color images, thanks to a lenticular screen [3]
- Free view point video (FVV) [4]: it enables users to view a 3D scene by freely changing their viewpoints.

In all cases, the quality of experience (QoE) strongly depends on the availability of a high number of views, closely arranged around or in front of the scene. In a high definition (HD) context, the acquisition and further the transmission of such a number of view streams become rapidly a blocking point for the development of these kinds of applications. Therefore, the multi-view plus depth (MVD) format allows capturing and transmitting to the 3D renderer a reduced number of view in the form of N video streams and the N corresponding per-pixel depth streams. At the receiver side, the missing intermediate views are synthesised thanks to a depth-image-based rendering (DIBR) algorithm [5]. In the development of multi-view-based immersive applications, the processing of the depth information, from the extraction to the exploitation at client side, is a key factor and has given rise to many research works. Especially, the depth map estimation remains a highly challenging problem. Stereo matching has been a very active research in the last few years. Basically, the depth map estimation process can be subdivided into four steps [6]: matching cost computation, cost aggregation, disparity computation and disparity refinement.

The disparity computation is probably the processing step which has been mostly explored in the literature and the step for which a important variety of computation methods has been proposed. One among others is the graph cut approach which is integrated in the reference depth map estimation tool used by the MPEG consortium [7]. Compared to the local winner-take-all strategy, this advanced disparity computation method produces depth maps of better quality but at the price of huge increase of the processing time. On the other hand, a semi-global matching (SGM) approach proposed in [8] allows to achieve good performances while keeping an acceptable complexity. Hence, by applying this basic algorithm it is possible to generate coarse disparity maps in real-time that can be used for monitoring purposes. In this chapter we intentionally focus on local disparity map computation methods in order to stay as close as possible to real-time or near real-time performances. Hence, even if they allow achieving more accurate depth maps, the global methods are not considered here and the preferred approach is to better exploit the availability of several views (more than two) in order to optimize the 3D rendering and especially the synthesis of intermediary views.

In the specific context of multi-view rendering, the quality of depth maps is very important to aid synthesis of views with high visual quality at arbitrary viewing angles. Fast depth extraction methods that operate in real-time (or near real-time) and are mainly based on disparity estimation yield in spatially and temporally inconsistent depth values at several regions. These inconsistencies often lead to lower quality intermediate view synthesis with visible artifacts. Furthermore, the video object boundaries in the depth maps and the corresponding regions in the

color images may not be well aligned. This would also result in synthesis artifacts at the object borders. To overcome these inherent limitations of disparity-estimation based depth map generation process, an enhancement processing can be applied by improving the temporal and multi-view coherence of the already estimated depth maps. In [9], the authors show that increasing consistency of the depth maps of the different views, the video coding scheme (multi-view plus depth) will achieve better performances.

In the remaining, we will firstly describe the multi-view acquisition system and the different post-processing stages which aim at achieving optimized multi-view content at rendering side. Next, the proposed algorithms are benchmarked and experimental results are provided. Finally conclusions are drawn, highlighting the key aspects of this study.

4.2 Acquisition and Post-processing

4.2.1 Devices Architecture

For the purpose of 3D video capturing, a multi-view system made of four industrial cameras, depicted in Fig. 4.1, has been built. The main features, concerning the multi-view aspects, of this system can be summarized as follows:

- Transportable rig,
- Linear (1-D Parallel) camera arrangement,
- Focal length: 9 mm,
- Inter-camera distances: 10 cm,
- Video format: YUV4:2:0, HD 1080p/25fps.

In Fig. 4.1, the scheme of the rig with the camera positions is provided and will be used in the sequel as the reference naming when describing the different multi-view process: view 1 refers to camera C_1, view 2 refers to camera C_2, ...

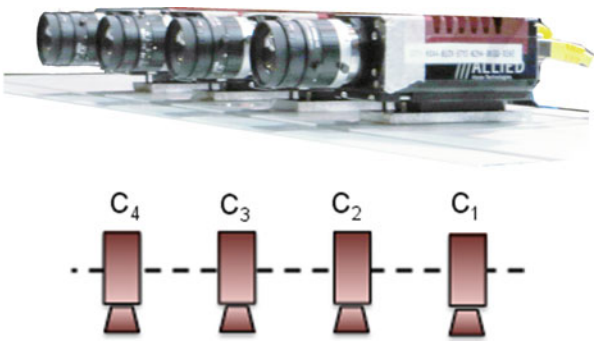


Fig. 4.1 Transportable multi-camera system and camera naming

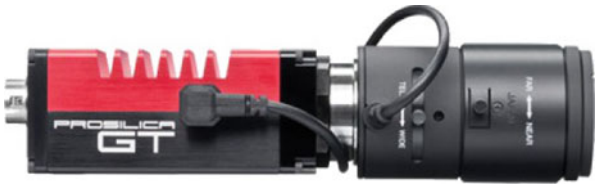


Fig. 4.2 Industrial camera model (AVT, Prosilica GT1910C)

Fig. 4.3 Camera trigger (Gradasoft CC320)

4.2.1.1 Cameras and Trigger

In terms of image resolution, the requirement was to be able achieving HD quality: 1080p, 25 fps. Moreover, the choice of the cameras was motivated by the possibility to easily transport the entire system and to build it rapidly with medium-cost equipments excluding high-end cameras from the broadcast world. Thus, in order to fit with a limited budget, the industrial camera depicted in Fig. 4.2 has been chosen. Another important feature of these cameras is to allow user getting directly the raw stream (Bayer) instead of RGB or YUV. This is specially interesting in terms of bandwidth, a Bayer frame being composed of 1,920 $\times$ 1,080 Bytes while an RGB frame will need 3 Bytes for each pixel. Thus, a bandwidth of only 50 MB/s is required for each camera (instead of 150 MB/s in RGB). In order to synchronize the cameras on the same capturing instant, the Prosilica accepts an external trigger. The choice of this external synchronizer has been focused on the Gardasoft CC320 (see Fig. 4.3) which allows triggering up to eight outputs. Finally, the last feature provided by this device is the GigE Vision interface that provides a bandwidth up to 125 MB/s using the Ethernet protocol.

4.2.1.2 Acquisition and Storage

One of the most challenging aspects of HD multi-view capturing system is its ability to acquire all the frames continuously, without any discarding. Frame drooping is definitely not acceptable due to the strict synchronization constraint

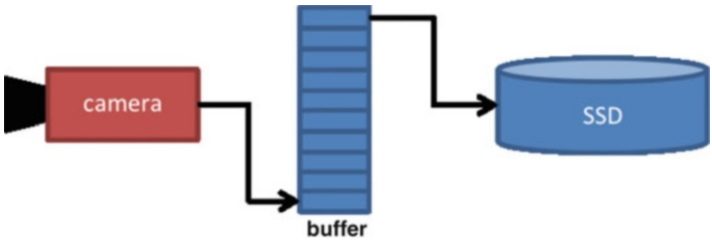


Fig. 4.4 Circular buffer principle

required for stereo rendering. To respect this constraint, the capture workstation has been equipped with the following hardware:

- Network adapters: composed of 4 Gigabits Ethernet port in order to get the stream of each camera separately.
- Hard drives: composed of two SSD hard drives able to write data up to 250 MB/s. Two camera streams are written on each device.

The writing rates of the hard drives are estimated statistically so that for a short period, this rate could be lower than the required bandwidth. To fix this issue, a software circular buffer has been used. In principle, it allows storing up to 50 frames into the Random Access Memory of the server. The concept is illustrated in Fig. 4.4. After the real time video content acquisition, a second step consists in post-processing the content in order to achieve an exploitable content in terms of 3D rendering. Classically, three main steps are needed: color space conversions, geometrical rectification, and color equalization between views.

4.2.1.3 Color Space Conversion

Due to bandwidth constraints previously mentioned, the streams are recorded in 8 bits Bayer format. Thus, the first post-processing step consists in converting these raw data to the simple red green blue format. To achieve this transformation, the method proposed by Sakamoto in [1] has been implemented. In our case, this conversion is not supposed to be done in real-time and can be easily implemented and executed on a CPU. Nevertheless, one can note that this kind of image processing is well adapted to be executed on a GPU, allowing to achieve real-time performances.

4.2.1.4 Geometrical Rectification

Once the converted RGB multi-view videos are available, errors due to the mechanical imperfections of a real system must be compensated. These corrections mainly aim at canceling lenses imperfection and geometrical shift of the camera sensor.

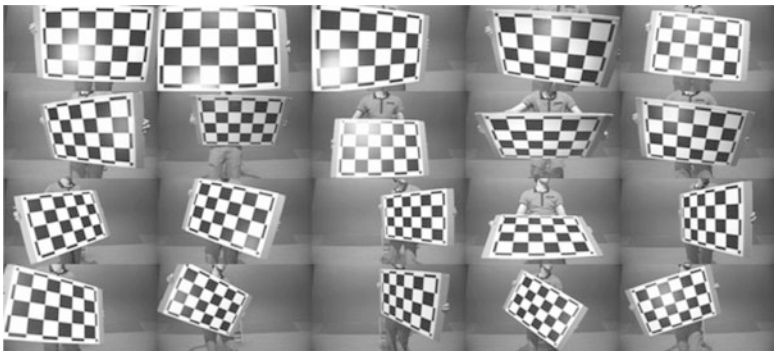


Fig. 4.5 Chessboard calibration samples

To achieve this geometrical enhancement, intrinsic and extrinsic parameters of the four cameras must be extracted. The intrinsic parameters correspond to the lenses distortion (radial and tangential) and the focal value. The extrinsic parameters are related to the relative position of the camera's sensor in the real world. This position is represented by a 3×3 rotation matrix and a three components translation vector. The extraction of these parameters is done using the well-known chessboard calibration method. Around 15–20 different views must be acquired to have an accurate calibration (see Fig. 4.5). In [10], Kang et al. present a simple and efficient method that extends the stereo image rectification presented by Fusiello in [11] to parallel multi-view images. The algorithm is divided into two main stages. Firstly, thanks to the intrinsic parameters, the lenses distortions are corrected. Next, the alignment of the four sensor plane is obtained, thanks to the extrinsic parameters (rotation and translation). This second step is depicted in Fig. 4.6. The last step of the rectification consists in cropping and resizing the frames in order to keep only the valid pixels of all the four cameras. In principle, this process aims at finding the valid common area between the views and to resize each camera view accordingly.

4.2.1.5 Color Equalization

Variations in camera parameters, different illumination conditions, or changes in viewpoint often cause changes in the color value of corresponding regions in two images of a scene. Those variations can lead to major problems during 3D processing. Therefore, a color calibration must be fulfilled to make the 3D algorithms more accurate. Its goal is to set two images or more to the same color distribution. In our case, the method proposed in [12] has been used as a basis and adapted in the multi-view context. This method can be divided into two sub-modules detailed in the sequel.

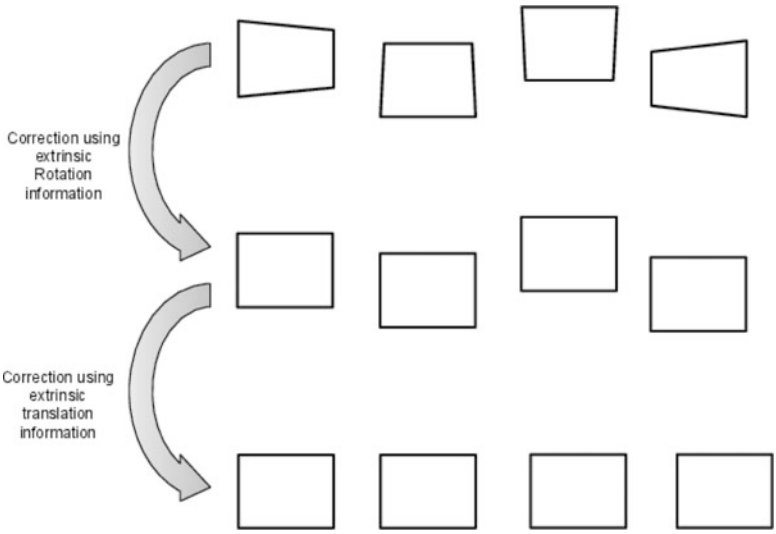


Fig. 4.6 Camera plan alignment

Color Space Transformation

A key requirement of the chosen color calibration algorithm is to work in an uncorrelated color space. In a classical color space like RGB space, all the channels are dependent from the others preventing color transfer operations. An ideal color space for these techniques is orthogonal, without any correlation between the components. In [13], Ruderman et al. have developed a color space, called $l\ \alpha\ \beta$, which minimizes correlation between channels for many natural scenes. This space is based on data-driven human perception studies which assume that the human visual system is ideally suited for processing natural scenes

Statistics and Color Correction

The goal of the algorithm is to transfer the color distribution of a reference image to a source image. This is done by recalculating the means and the standard deviations along the three axes l, α, and β. These parameters are calculated separately for each channel, for both the source and the reference images. First, the mean values are subtracted from each component of the images:

$$\begin{aligned} l^* &= l - \mu_l \\ \alpha^* &= \alpha - \mu_\alpha \\ \beta^* &= \beta - \mu_\beta \end{aligned} \tag{4.1}$$

Fig. 4.7 Visual effect of the color and geometrical correction on cameras C_1 and C_2 (*top*: before rectification, *bottom*: after rectification)

Then, normalization factors are calculated from standard deviations and a scaling operation is applied as follows:

$$\begin{aligned} l' &= \frac{\sigma_t^l}{\sigma_s^l} l^* \\ \alpha' &= \frac{\sigma_t^\alpha}{\sigma_s^\alpha} \alpha^* \\ \beta' &= \frac{\sigma_t^\beta}{\sigma_s^\beta} \beta^* \end{aligned} \tag{4.2}$$

where t corresponds to the target image used as a reference and s refers to the source image to be adapted.

After this transformation l', α', and β' represent the $l\ \alpha\ \beta$ components of the source image with the same standard deviation as the reference image. Finally, the mean value of the reference image is added resulting to an image with the statistics of the reference.

In Fig. 4.7 the visual effects of the geometrical and color corrections are illustrated. Especially, thanks to the delimitation (white dashed line) between the ground and the green wall in the background, the horizontal alignment is highlighted. Concerning the color correction, it is worth noticing that the big difference in terms of colorimetry is mainly due to specific correction applied to compensate some limitations of the cameras' sensors rather than color equalization between views.

4.3 Depth Map Computing

4.3.1 Depth Map Extraction

Depth and disparity, respectively, noted Z and "Disp" (in pixel unit), are inversely related according to the following equation:

$$\text{Disp} = (f.T)/(Z.t_{\text{pixel}}) \tag{4.3}$$

where f is the focal distance, T the inter-camera distance, and t_{pixel} the width of one pixel of the camera sensor.

For simplicity, in the remaining of the chapter f is assumed to be given in pixel unit and thus $\text{Disp} = (f.T)/Z$

Accurate dense stereo matching is a very challenging task. Especially in the image areas containing occlusions, object boundaries or fine structure can finally appear blurred inside the depth map. Lower repetitive textures, illumination differences, and bad recording conditions can also make this task even more complicated. In the sequel, an algorithm based on SGM and mutual information method [8] is briefly described and used further as a general framework for the experiments. This depth computation method achieves good performances and the algorithm can be substantially accelerated, thanks to several possible optimizations. Especially, this algorithm can be executed in real time and can be used for monitoring purposes. The SGM stereo method relies on block matching of mutual information and the approximation of a global 2D smoothness constraint by combining many 1D constraints. The method can be divided into three steps that are described in the sequel.

4.3.1.1 Block Matching Cost Calculation

The goal of this step is to compute the difference between two areas in the left and right images to find the corresponding pixels. In order to find them, the left image is divided in blocks and each block is compared with a same size block in the right image that is moved in a defined range (called "Horopter") on the same line. A matching function computes "the cost" associated with each tested area. The cost is calculated as the absolute minimum difference of intensities between the tested blocks. Several parameters can be set during this step like the size of the matching window or the size of the blocks. All the costs are saved to be used in the next step of the algorithm.

4.3.1.2 Cost Aggregation

The basic block matching cost calculation is potentially ambiguous and wrong matches can have a lower cost than the correct ones due to a poor awareness of image content. Therefore an additional constraint is added that supports smoothness by penalizing changes of neighboring disparities. Thus the cost aggregation is based on a penalty-reward system and it is processed among eight paths from all direction around the block.

4.3.1.3 Disparity Computation

Based on the previous cost aggregation step, the disparity image is determined by selecting for each pixel p the disparity d that corresponds to the minimum calculated cost. To avoid problems brought about by not enough textured regions, a parameter named uniqueness constraint is introduced. It enables to check the consistency of a region and adapt the disparity computed in regards of it.

As previously mentioned, this three steps disparity map calculation framework is a basis for numerous fast block matching algorithms. In [14], the authors propose an optimized version of this kind of local disparity computation method. This study shows that it is possible to produce depth maps of high accuracy with a local bloc matching algorithm by applying advanced filtering processes which improve the spatio-temporal consistency. In addition, the accuracy of the map can be improved again by refining the decision process, with the use of an optimized reliability criterion, after the aggregated cost function (ACF) calculation. In Sect. 4.4, this algorithm will be used as a basis in order to test the multi-view refinement algorithm proposed in the next section.

4.3.2 Multi-view Depth Refinement

4.3.2.1 Edge Adaptive Median Filtering for Multi-view Depth Maps

This first stage of multi-view depth processing framework comprises three sub-stages as view warping, adaptive median filtering, and inverse view warping, as was proposed in [15]. First, considering N total number of viewpoints and according depth maps, all depth viewpoints are projected to the image coordinates of the center viewpoint $N/2$. This process is depicted in Fig. 4.8. The depth maps are transformed to the real world depths by:

$$D(x,y,t,n) = \left(\frac{d(x,y,t,n)}{255} \cdot \left(z_{\text{near}}^{-1} - z_{\text{far}}^{-1}\right) + z_{\text{far}}^{-1}\right)^{-1} \tag{4.4}$$

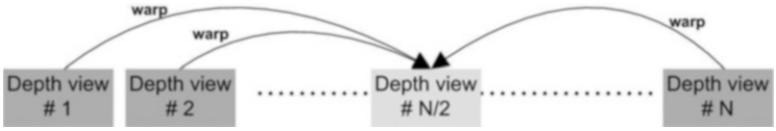


Fig. 4.8 N views projection to the image coordinates of the center view

where x and y depict the image coordinates, t denotes the time, n denotes the viewpoint number, z_{near} and z_{far} denote the smallest and largest depth in the scene (in physical units). D and d stand for the actual depth value and quantized depth value (in the range of 0–255) of a pixel, respectively. Afterwards, these transformed depth values are used to obtain the three-dimensional coordinates as:

$$(u, v, w) = R(n) \cdot A^{-1}(n) \cdot (x, y, 1) \cdot D(x, y, t, n) + T(n) \tag{4.5}$$

where the functions $A(n)$, $R(n)$, and $T(n)$ represent the intrinsic parameters, rotation, and translation, respectively, of the nth camera viewpoint. Then, all coordinates are projected back to the image coordinates of camera n_0, which is the center camera:

$$(u', v', w') = A(n_0) \cdot R^{-1}(n_0) \cdot \{(u, v, w) - T(n_0)\} \tag{4.6}$$

The projected coordinates (x', y') are expressed in a homogeneous two-dimensional form as $x' = u'/w'$ and $y' = v'/w'$. Finally, the warped depth map values are denoted as:

$$d'_{n_0}(x, y, t, n) = d(x', y', t, n) \tag{4.7}$$

The forward warping process is followed by the main processing cycle, i.e. the adaptive median filtering, incorporating all the warped depth map frames. The mentioned median filter is applied on a four-dimensional window, S. The shape of the window S is adaptive to edges and also the motion in the multi-view sequences. Hence, the adaptation factors are defined as: (1) the local variance of the depth values and (2) the local mean of the absolute difference between the luminance components of the two consecutive color video frames. They are denoted as $V(x, y, t)$ and $m(x, y, t)$, respectively. Both parameters are computed in a 5×5 spatial neighborhood $N_{t,n_0}(x, y)$ taken at the time instant t and viewpoint n_0 and are centered around the spatial location (x, y). Accordingly, the first parameter is computed as:

$$V(x, y, t) = \text{var}\left(D'_{n_0}\left(N_{t,n_0}(x, y)\right)\right) \tag{4.8}$$

And the second parameter is computed as:

$$m(x, y, t) = \text{mean}\left(\left|c\left(N_{t,n_0}(x, y) - c\left(N_{t-1,n_0}(x, y)\right)\right)\right|\right) \tag{4.9}$$

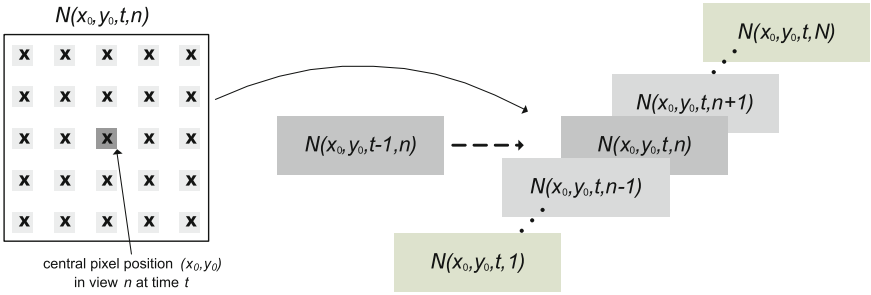


Fig. 4.9 Sample multi-dimensional window, when both the variance and motion factors are smaller than their respective threshold values

The two parameters are compared to two different threshold values Thr_v and Thr_m, respectively. For instance, if the variance is lower than the threshold (i.e., not an edge), the window includes all the 5 $\times$5 neighborhoods across all warped depth representations. Otherwise, the window is restricted only to the center pixel across all warped depth representations. Moreover, if no motion is detected, i.e. the computed m is lower than Thr_m, then the temporal coherence is enforced by including the locations at the previous time instant $(t-1)$ in the window. An example of the window, where both factors are smaller from their respective threshold values is given in Fig. 4.9 [15].

Following the construction of the adapted multi-dimensional window, the resulting depth values are obtained by median filtering on the same window as:

$$D'_{n_0}(x, y, t, n) = \text{median}[D'_{n_0}(N(x, y, t))] \tag{4.10}$$

For all cameras. It should be noted that the inter-view coherence is achieved in this case by using the same resulting real world depth value for all views $n = 1, \ldots, N$. Following this computation, the filtered real world depth values are mapped back to the luminance values ranging from 0 to 255, and the inverse view warping takes place to map back the filtered depth luminance values to their original viewpoint's image coordinates. The occluded depth pixels in the image coordinates of the center viewpoint are processed in subsequent stages, by setting the target viewpoint as the cameras numbered $\lceil N/2 \rceil \pm 1$, $\lceil N/2 \rceil \pm 2, \ldots$ instead of $\lceil N/2 \rceil$.

4.3.2.2 Color Texture/Depth Edge Alignment on the Resultant Depth Maps

After the first processing stage, in order to make the resultant depth maps easier to compress, the median filtered depth values are smoothed using a joint-trilateral filter, which incorporates the closeness of depth values, as well as the similarity of both the color and depth map edges. In each depth viewpoint, for each pixel to be processed, a window of $2w \times 2w$ is formed centered at that particular depth pixel.

Subsequently, in these kernels of $2w \times 2w$ (denoted by Ω), the filtered depth value is computed as:

$$D''_p = \frac{\sum_{q \in \Omega} \text{coeff}_{pq} \cdot D'_q}{\sum_{q \in \Omega} \text{coeff}_{pq}} \tag{4.11}$$

where q denotes a pixel within the kernel, p is the center pixel to be processed and D'_q the depth value of pixel point q obtained from (4.10). The "coeff" is a multiplication of three different factors, namely the closeness in pixel, similarity in depth value, and the similarity in color texture value:

$$\text{coeff}_{pq} = c(p,q) \cdot s_{\text{depth}}(p,q) \cdot s_{\text{color}}(p,q) \tag{4.12}$$

The filters related to these three factors are considered as Gaussian filters centered at point p. Accordingly, these individual factors are denoted as:

$$\begin{aligned} c(p,q) &= \exp\left(-\frac{1}{2}(p-q)^2/\sigma_c^2\right) \\ s_{\text{depth}}(p,q) &= \exp\left(-\frac{1}{2}(d_p - d_q)^2/\sigma_{s_{\text{depth}}}^2\right) \\ s_{\text{color}}(p,q) &= \exp\left(-\frac{1}{2}(I_p - I_q)^2/\sigma_{s_{\text{color}}}^2\right) \end{aligned} \tag{4.13}$$

where d represents the depth values of pixel points p and q, and I represents the corresponding color texture luminance values. The standard deviations are calculated for each kernel, such that the group of pixels within the kernel (of size $2w \times 2w$) fall into the 95 % confidence interval in the Gaussian distribution [9]. Hence, all the standard deviation parameters within these three factors are equal and calculated as:

$$\sigma_c = \sigma_{s_{\text{depth}}} = \sigma_{s_{\text{color}}} = w/2 \tag{4.14}$$

Figure 4.10 depicts a visual example from a depth frame of three neighbor viewpoints (from *Pantomime* sequence), before and after the application of the two stages of processing. From Fig. 4.10, it can be seen that many obvious textural differences existing among the depth frames of the same time instant are removed as a result of the application of the processing stage. Furthermore, many textural gradients are also smoothened while preserving and correcting the depth edge transitions in the second stage of the processing.

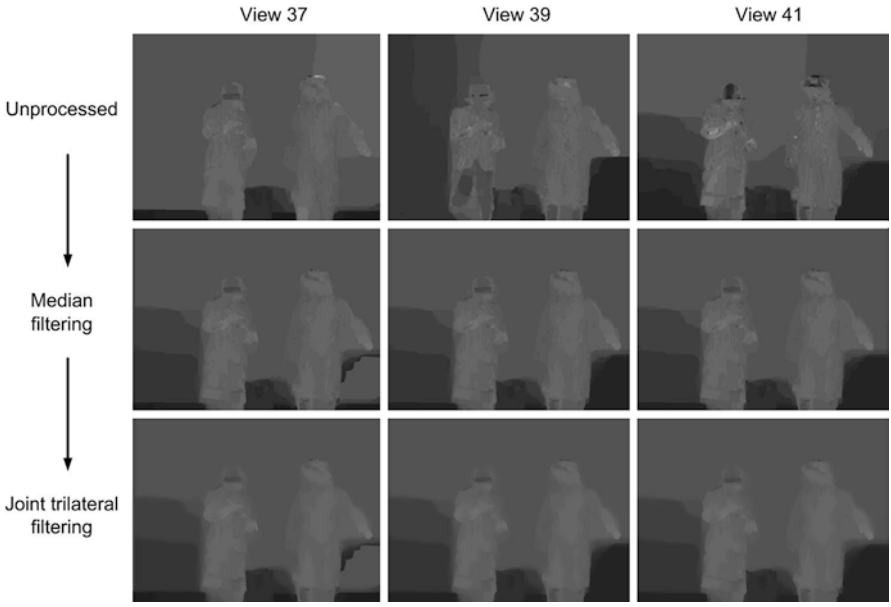


Fig. 4.10 State of the multi-view depth maps before and after the processing stages

4.4 Experimental Results

In order to validate the multi-view capturing system and the disparity map extraction methods presented previously, different contents have been captured in the following conditions:

- Resolutions: HD 1080p/25fps,
- Pre-processing: inter-view alignment and color equalization,
- Inter-camera distances: $T = 10$ cm,

The simulations results presented in this section are obtained from the processing of streams coming from two shooting scenario:

- "Musicians": three persons playing music statically
- "Actors": three persons talking and moving on a stage.

For each sequence, a sample of 2 s (50 frames) has been extracted and further used for the simulations.

In order to evaluate the quality of the disparity maps, a classical approach consists in calculating the objective quality metric (e.g., PSNR) of a depth image-based rendered view. In the sequel, we will focus on cameras C_1, C_2, and C_3 (see Fig. 4.1) and the disparity map will be calculated between cameras C_1 and C_2 or cameras C_2 and C_3. Then the views corresponding to cameras C_1 and C_3 are reconstructed from the view corresponding to C_2 and the disparity maps.

4.4.1 Depth Map Extraction

In this study, the development of a new stereo depth map extraction algorithm, beyond state of the art in term of depth accuracy, is not targeted. On the other hand, fast and practical algorithms have been used in order to provide depth of rather high quality that are enhanced next by exploiting the multi-view aspects. Thus, in this section, we provide visual results of the kind of disparity one can produce in real-time or near real-time conditions with the two disparity extraction algorithms presented before. Depth rendered views are also provided with the corresponding quality metrics which are more intensively used further (Sect. 4.4.2) when evaluating the coding performances.

4.4.1.1 Semi-Global Block Matching

In Fig. 4.11, visual results of the semi-global block matching approach are provided. The top left picture corresponds to the first frame of the original stream (after geometrical and color correction) coming from camera C_3. The top right picture is a gray scale representation of the disparity between camera C_3 and camera C_2. This disparity is the result of the SGBM algorithm, followed by a Gaussian blurring filter applied with a 51×51 kernel. The disparity map directly provided by the SGBM algorithm is of good accuracy, but when using it for DIBR without more processing, many occluded pixels appear in the synthesized view. On the other hand, applying the blurring filter on the coarse disparity map slightly decreases the accuracy of the depth but allows reconstructing more pixels. Hence, on the bottom right picture, which represents the reconstructed view 3, without enhancement processing, we can see the occluded pixels in black. This amount of occluded pixels ($\sim$5 %) remains quite low and allows the appliance of inpainting technique [16, 17] in order to reconstruct the entire video frame. The entire reconstructed frame of view 3, after inpainting, is represented on the bottom left of Fig. 4.11. The quality of this picture measured by the PSNR ($\sim$33 dB) is definitely good. However, strong object deformations can be observed, especially for the microphone stands in the foreground. Even if it is not really visible due to the black areas in the bottom right picture, these deformations are already present before inpainting the image and are directly produced by the DIBR. Thus, these artifacts visible on slim objects clearly illustrate the main drawback of the blurring filter (loss of high frequencies) applied on the depth map. These first simulations results highlight the need of a very accurate disparity map in order to allow high quality view synthesis.

Fig. 4.11 Disparity map and view synthesis with SGBM (*top left*: original view 3, *top right*: disparity map, *bottom left*: synthetised view 3 with inpainting (32. 97 dB), *bottom right*: synthetised view 3 with 4 % of occluded pixels)

4.4.1.2 Accurate Local Block Matching

In a second time, the accurate local block matching algorithm proposed in [14] has been tested in order to generate more accurate disparity maps while executing in near real-time (∼11 frames per second), thanks to a GPU-based implementation. Regarding the disparity map (top right) obtains with SGBM algorithm, one can clearly visually appreciate the improvement brought by the accurate block matching algorithm. Despite an important level of noise on the background, the different objects of the scene are remarkably well segmented. In this case, any blurring algorithm has been applied and the disparity map has been used directly from the block matching in order to render the bottom right reconstructed picture with occlusions. The number of occluded pixels is slightly the same than what was obtained with SGBM algorithm and the PSNR after inpainting (bottom left) is very close to 33 dB as well. However, the microphone stands are significantly better reconstructed with this method, showing the benefit of this accurate block matching approach (Fig. 4.12).

In addition, for this second method, it is important to note that the disparity map has been obtained after merging the two disparity maps: one computed from C_2 to C_1 and the other computed from C_2 to C_3, leading to lower performances than the ones obtained with only the disparity from C_2 to C_3. Indeed, in a MVD transmission scheme, it is classically expected to transmit only one disparity or depth map per view. Since in a multi-view system the disparity can be computed from both right and left sides, this implies to efficiently merge the two disparities without

Fig. 4.12 Disparity map and view synthesis with accurate local block matching (*top left*: original view 3, *top right*: disparity map, *bottom left*: synthetised view 3 with inpainting (32. 81 dB), *bottom right*: synthetised view 3 with 5 % of occluded pixels)

penalizing the right or the left reconstructed frames quality at rendering. In the next section, regarding this information coding challenge, different compression schemes are proposed and evaluated.

4.4.2 Depth Map Enhancement and Coding Performances

In this section, it is assumed that the disparity maps are transmitted through a limited bandwidth channel. The objective here is to evaluate different coding schemes of the disparity maps without considering any other compression methods for the corresponding views. Thus, in order to not introduce any other bias in the quality of the reconstructed frames, the view which is used as an input for the DIBR does not suffer from any kind of lossy coding artifacts. Obviously, in a practical use case the four views are compressed to fit with bandwidth constraints.

Taking into account of the disparity map transmission, different coding schemes can be considered. As previously explained, when considered a four camera multi-view system, six different disparity maps can be theoretically extracted: from C_1 to C_2 (Disp_{12}), from C_2 to C_1 (Disp_{21}), from C_2 to C_3 (Disp_{23}), from C_3 to C_2 (Disp_{32}), from C_3 to C_4 (Disp_{34}), and from C_4 to C_3 (Disp_{43}). Hence, the first transmission scheme consists in transmitting these six disparity maps without any merging process. Although this scenario is bandwidth consuming, it is even possible to reduce the bitrate by using an adapted coding format like MPEG-4 multi view coding (MVC) [18]. For instance, by using MVC, the two disparity maps: Disp_{21}

and Disp_{23} associated with view 3 are encoded in the same video elementary stream. In the sequel this scenario is referred to as the "two disparity maps" scenario and the video elementary stream is obtained, thanks to JM reference software [19] which supports MVC encoding.

On the other hand, instead of coding the two disparity map streams, a merging process can be applied on the corresponding depth maps. A simple approach consists in computing the mean depth obtained from Disp_{21} and Disp_{23} before encoding and transmission:

$$Z_2 = (Z_{21} + Z_{23})/2 \tag{4.15}$$

where $Z_{21} = f.T/\text{Disp}_{21}$ and $Z_{23} = f.T/\text{Disp}_{23}$.

After simplification, we can write:

$$\frac{Z_2}{f.T} = \frac{1}{2}\left(\frac{1}{\text{Disp}_{21}} + \frac{1}{\text{Disp}_{23}}\right) \tag{4.16}$$

Let's call $\text{Disp}_2 = f.T/Z_2$ the new disparity obtained after the merging process:

$$\text{Disp}_2 = 2.\left(\frac{\text{Disp}_{21}.\text{Disp}_{23}}{\text{Disp}_{21} + \text{Disp}_{23}}\right) \tag{4.17}$$

In the sequel, this approach is referred to as the "simple merge" scenario and the merged video stream Disp_2 is then encoded in MPEG-4 AVC format with the JM encoder. Finally, the third approach referred to as "enhanced disparity map" consists in merging the redundant disparity maps during the wrapping/inverse wrapping process described in (4.3.2) and which can be seen as an enhanced merging process regarding the merging described by (4.17).

The graphs presented in Figs. 4.13 and 4.14 show the multi-view depth map coding performances for "Musicians" and "Actors" sequences, respectively, with regard to the view synthesis quality (using PSNR metric). The horizontal axis shows the disparity map coding bitrate and the vertical axis indicates the PSNR values of the synthesized views: view 3 on top and view 1 on the bottom. As described in (4.4.1), a simple DIBR without post-processing (e.g., inpainting) generates images with occluded pixels. In the PSNR computing, these black pixels introduce an important bias, leading to unexploitable results. In order to avoid this drawback and for simplification purposes, these occluded pixels are excluded from the set of pixels used for PSNR calculation. By doing this, an unfair factor is potentially introduced when comparing the performances. However, by observing the percentage of reconstructed pixels (not occluded) provided in Figs. 4.15 and 4.16, one can notice that these values are very close for each evaluated approach. In all cases, the value never falls below 94 % and never exceeds 97 %, while the maximum distance between two scenarios at a given bitrate remains below 1 %. In addition, without going further in the analysis of the results, it is interesting to observe that in a general manner the number of reconstructed pixels

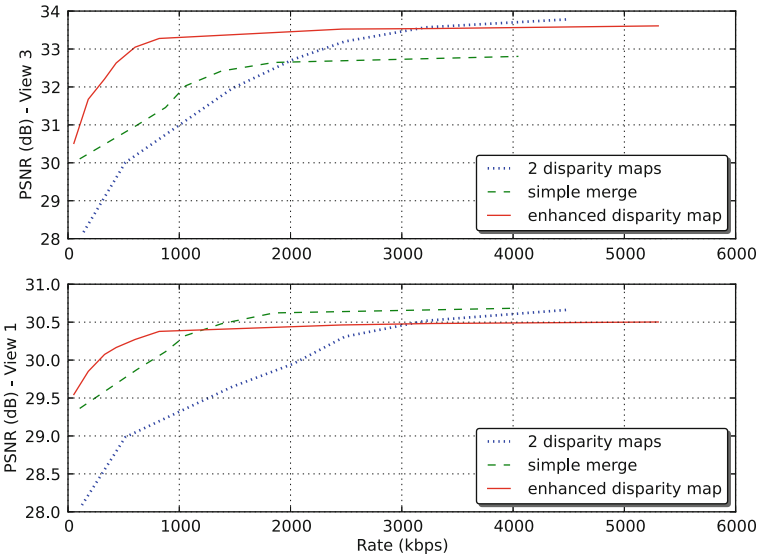


Fig. 4.13 Rate-distortion curves comparison, sequence "Musicians"

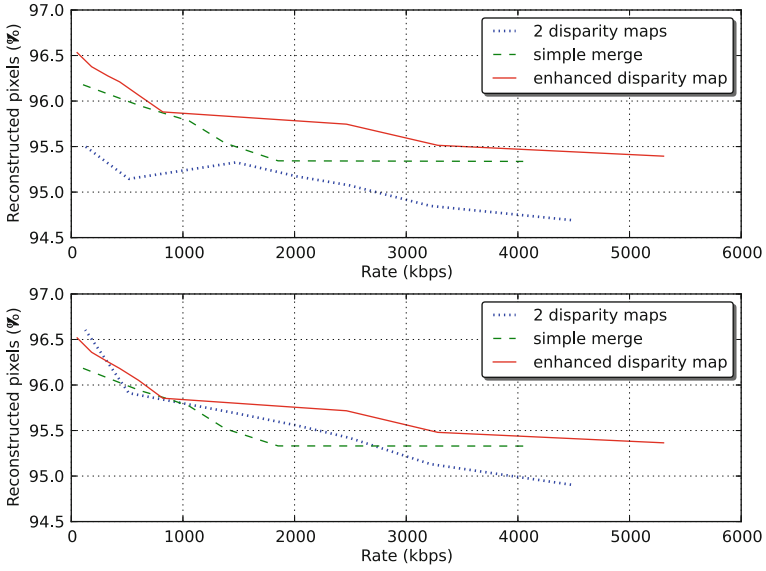


Fig. 4.14 Percentage of reconstructed pixels, sequence "Musicians" (*top*: view 3, *bottom*: view 1)

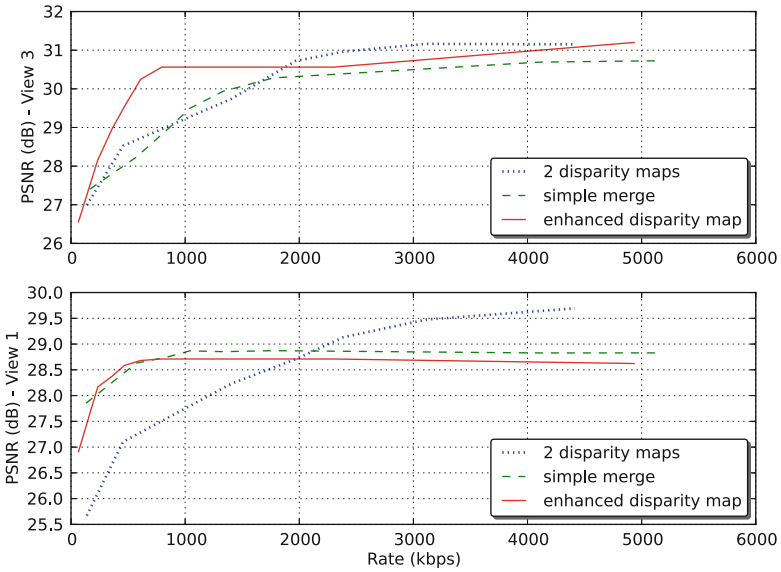


Fig. 4.15 Rate-distortion curves comparison, sequence "Actors"

is more important for the approaches where the measured PSNR is higher as well, which validates our comparison method in terms of fairness.

Figures 4.13 and 4.14 show that in average, the "merged" approaches (one disparity map per view) achieve better performance than the "two disparity maps" approach. In almost practical bitrate ranges, the "simple merge" approach outperforms the method based on two maps. Indeed, the averaging between depth results to a better spatial consistency of the maps and then to better coding performances. For the very low bitrates, the "enhanced disparity map" significantly outperforms the two others approaches. Especially, around 500 kbps, a gain higher than 1. 5 dB (up to 3 dB) is usually reached. These results demonstrate the efficiency of the proposed algorithm for the improvement of the spatial, temporal, and multi-view consistency of the maps leading to maximized compression performances.

Regarding Figs. 4.15 and 4.16, it is interesting to note that the percentage of reconstructed pixels is monotonically decreasing functions with the bitrate for the merged approaches whereas increasing the compression sometimes allows reconstructing more pixels with the "two disparity maps" approach. Actually, without any consistency enhancement after disparity map extraction, the video compression process somehow acts as a lowpass blurring filter, increasing the number of reconstructed pixels, as observed in (4.4.1), while degrading the video quality.

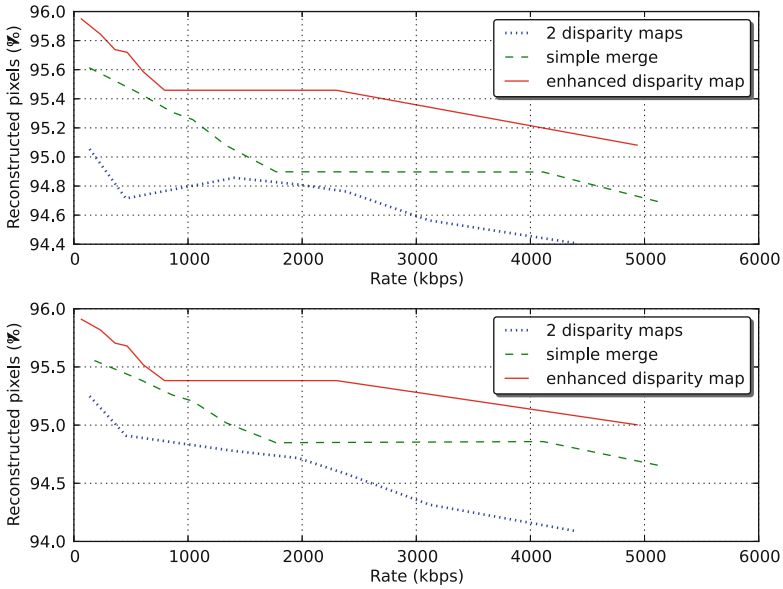


Fig. 4.16 Percentage of reconstructed pixels, sequence "Actors" (*top*: view 3, bottom: view 1)

Finally, Fig. 4.17 provides a visual comparison of a reconstructed video frame (at 450 kbps), between the "two disparity maps" and the "enhanced disparity map" approaches. This figure clearly shows the benefit of the enhanced method in terms of rendering artifacts. Especially, the faces of the two actress are strongly deformed or occluded on the borders in the central pictures. In addition, even if some artifacts are visible on the stepladder on the picture obtained with the enhanced maps (bottom), the general shape of the object is really much preserved.

4.5 Conclusion

This chapter provides a complete framework for multi-view acquisition, with a deep focus on depth map processing, in order to provide high quality views and depth maps streams to the encoding stage and to allow optimized intermediate view synthesis at rendering side. The capturing devices are dimensioned in order to perform real-time 1080p@25fps HD acquisition. As presented in the experimental results section, coarse disparity maps with acceptable quality can be extracted by implementing a fast version of the SGBM algorithm. While executing very fast, a more complex and accurate block matching-based method allows achieving disparity maps with significant improvements. Different coding schemes, for the

Fig. 4.17 Visual comparison of reconstructed view 3 with disparity map encoded at 450 kbps, sequence "Actors" (*top*: original frame, *middle*: two disparity maps per view, *bottom*: enhanced disparity map per view)

transmission of the disparity maps are proposed. Finally, it has been demonstrated that applying a depth map refinement algorithm allows enhancing the depth images and then significantly increasing the quality of the synthetised views in a constrained bitrate environment.

Acknowledgments This work was supported by the ROMEO project (grant number: 287896), which was funded by the EC FP7 ICT collaborative research program.

References

1. Smolic A (2011) 3D video and free viewpoint video-from capture to display. Pattern Recognit 44(9):1958–1968. [Online] http://dx.doi.org/10.1016/j.patcog.2010.09.005
2. Kubota A, Smolic A, Magnor M, Tanimoto M, Chen T, Zhang C (2007) Multiview imaging and 3DTV. IEEE Signal Process Mag 24(6):10–21
3. Matusik W, Pfister H (2004) 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. ACM Trans Graph 23(3):814–824. [Online] http://doi.acm.org/10.1145/1015706.1015805
4. Tanimoto M (2010) FTV (Free-viewpoint TV). In: 17th IEEE international conference on image processing (ICIP), September 2010, pp 2393–2396. [Online] http://dx.doi.org/10.1109/ICIP.2010.5652084
5. Müller K, Smolic A, Dix K, Merkle P, Kauff P, Wiegand T (2008) View synthesis for advanced 3D video systems. EURASIP J Image Video Process 2008:1–11. [Online] http://dx.doi.org/10.1155/2008/438148
6. Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vis 47(1–3):7–42. [Online] http://dx.doi.org/10.1023/A:1014573219977
7. Tanimoto M, Fujii T, Panahpour M, Wildeboer M (2009) Depth estimation reference software DERS 5.0. ISO/IEC JTC1/SC29/WG11, Technical Report M16923
8. Hirschmuller H (2008) Stereo processing by semiglobal matching and mutual information. IEEE Trans Pattern Anal Mach Intell 30(2):328 –341
9. De Silva D, Fernando W, Kodikaraarachchi H, Worrall S, Kondoz A (2010) Adaptive sharpening of depth maps for 3D-TV. Electron Lett 46(23):1546–1548
10. Kang Y-S, Lee C, Ho Y-S (2008) An efficient rectification algorithm for multi-view images in parallel camera array. In: 3DTV Conference: the true vision - capture, transmission and display of 3D video, May 2008, pp 61–64. [Online] http://dx.doi.org/10.1109/3DTV.2008.4547808
11. Fusiello A, Trucco E, Verri A (2000) A compact algorithm for rectification of stereo pairs. Mach Vis Appl 12(1):16–22. [Online] http://dx.doi.org/10.1007/s001380050003
12. Reinhard E, Ashikhmin M, Gooch B, Shirley P (2001) Color transfer between images. IEEE Comput Graph Appl 21(5):34–41. [Online] http://dx.doi.org/10.1109/38.946629
13. Ruderman DL, Cronin TW, Chiao C-C (1998) Statistics of cone responses to natural images: implications for visual coding. J Opt Soc Am A 15(8):2036–2045. [Online] http://josaa.osa.org/abstract.cfm?URI=josaa-15-8-2036
14. Drazic V, Sabater N (2012) A precise real-time stereo algorithm. In: Proceedings of the 27th conference on image and vision computing New Zealand, ser. IVCNZ '12. Association for Computing Machinery, New York, pp. 138–143. [Online] http://doi.acm.org/10.1145/2425836.2425867
15. Ekmekcioglu E, Velisavljevic V, Worrall S (2011) Content adaptive enhancement of multi-view depth maps for free viewpoint video. IEEE J Sel Top Signal Process 5(2):352–361

16. Bradski G, Kaehler A (2008) Learning OpenCV: computer vision with the OpenCV library. O'Reilly, Sebastopol
17. Telea A (2004) An image inpainting technique based on the fast marching method. J Graph Tools 9(1):23–34. [Online] http://www.tandfonline.com/doi/abs/10.1080/10867651.2004.10487596
18. Vetro A, Wiegand T, Sullivan G (2011) Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. Proc IEEE 99(4):626–642
19. H.264/AVC JM Reference Software (2008) Joint video team (JVT) of ISO/IEC MPEG & ITU-T VCEG. [Online] http://iphome.hhi.de/suehring/tml/, August 2008

Chapter 5
Object-Based Spatial Audio: Concept, Advantages, and Challenges

Chungeun Kim

Abstract One of the primary objectives of modern audiovisual media creation and reproduction techniques is realistic perception of the delivered contents by the consumer. Spatial audio-related techniques in general attempt to deliver the impression of an auditory scene where the listener can perceive the spatial distribution of the sound sources as if he/she were in the actual scene. Advances in spatial audio capturing and rendering techniques have led to a new concept of delivering audio which does not only aim to present to the listener a realistic auditory scene just as captured but also gives more control over the delivered auditory scene to the producer and/or the listener. This is made possible by being able to control the attributes of individual sound objects that appear in the delivered scene. In this section, this so-called "object-based" approach is introduced, with its key features that distinguish it from the conventional spatial audio production and delivery techniques. The related applications and technologies will also be introduced, and the limitations and challenges that this approach is facing will follow.

5.1 What Is Object-Based Audio?

As briefly mentioned above, object-based audio fundamentally aims to record or capture the sound objects instead of audio channels which correspond to all the sound coming from given directions. The information necessary to place the individual objects in the rendered auditory scene as intended is also recorded or created. At the rendering side, this information is applied to the processing of each audio object signal such that the object sounds as if it were at the corresponding position, towards the corresponding direction. Attributes such as sound level, position, and orientation

C. Kim (✉)
I-Lab Multimedia and DSP Research Group, Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, UK
e-mail: chungeun.kim@surrey.ac.uk

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_5, © Springer Science+Business Media New York 2014

are therefore essential in this auditory scene information. The audio rendering system operates as a mixer which interprets the scene information and reproduces the scene through the given number of transducer feeds (headphones or multichannel loudspeakers) as mixtures of the objects. The placement and mixing of the objects can be achieved by applying perceptual cues to the source signals, such as interaural time and level differences (ITD and ILD) and head-related transfer function (HRTF) for binaural or conventional multichannel surround systems [1,3], or by physically generating the total sound field within a given listening area corresponding to the captured scene, in the case of ambisonic [8] or wave field synthesis (WFS) [2] systems. This implies that object-based audio is fundamentally independent of the rendering configuration—any can be used if the scene information is interpreted and used correctly for processing.

5.2 Differences from Conventional Channel-Based Approach

For distinction from the term object-based, the conventional method currently being used widely for spatial audio production can be called as "channel-based," since it captures and delivers audio signals for dedicated channels as mixtures of objects. Figure 5.1 conceptually describes and compares the typical operation from capturing to rendering of spatial audio in conventional channel-based approach to that in the object-based approach, for a 5.1-channel loudspeaker system as an example.

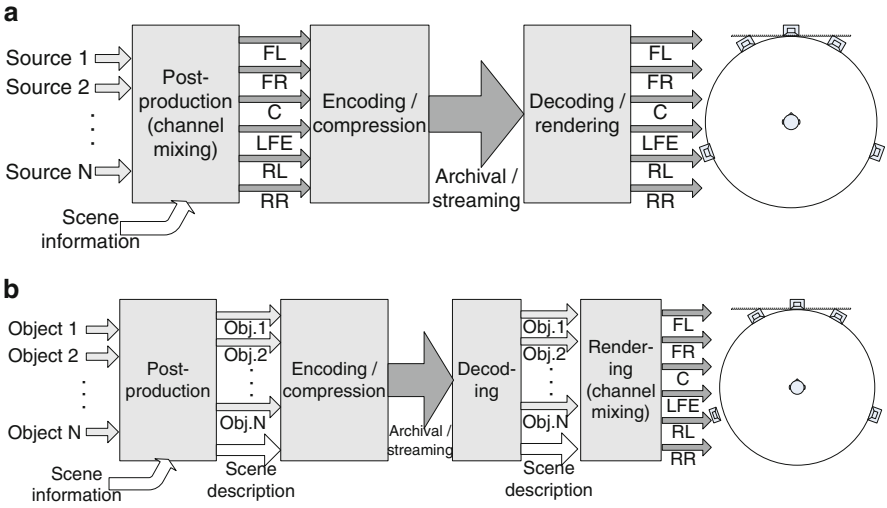


Fig. 5.1 Typical processing flow of audio signals for 5.1-channel loudspeaker configuration: (**a**) using conventional channel-based approach and (**b**) using object-based approach

In the channel-based chain, each input (denoted as source) typically corresponds to a sound object or sometimes a mixture of objects or various acoustic "effects" without actual objects (such as reverberation). The postproduction module operates mainly as a mixer that places the incoming sources at the intended positions, based on the scene information at the capturing site, in the auditory scene represented by the 5.1-channel loudspeakers. The output from this module usually has the form of a multichannel interleaved audio stream. It can be encoded further through compression, depending on the application, as an archived medium or as a stream towards a remote renderer. The decoder and/or renderer simply retrieve the produced audio channels and feed them to the corresponding loudspeakers.

On the other hand, in the object-based approach, the inputs mostly represent individual sound objects. The postproduction module prepares the objects along with the scene description, generated from the information of the captured scene. This needs to include all the detailed information as parameters required for correct rendering, such as the level, position, orientation, reverberation-related parameters (e.g., delay time and amount), and other spatial parameters (related to width and envelopment), that can vary with time. It also needs to be organized in a format that can be interpretable at the rendering side. Encoding and decoding processes can be introduced as in the channel-based approach, suitable for multi-object signals, to save the data storage or transmission bandwidth. The input to the rendering module is the same as the postproduction output. Compared to the channel-based approach, it is at this stage that the mixing and placing of objects happens, corresponding to the scene description. The processor inside the rendering module will produce the right feed to each loudspeaker channel using the scene description parameters and the object signals. Depending on the application, the end user at the rendering side can be given an additional control over the objects—that is, the user can adjust the scene description before the final rendering and thus the object distribution or characteristics in the auditory scene.

5.3 Advantages of Object-Based Approach

One of the main advantages of the object-based spatial audio over the channel-based approach is the enhanced control over individual objects. As mentioned already in the previous section, the object-based approach can give control over the scene description and therefore the individual objects both to the producer and to the end user, whereas in the channel-based approach, the detailed editing of individual sound objects is possible only by the producer. This feature is useful in applications where the user can interact with the auditory scene. For example, in video games, virtual reality (VR) applications or augmented reality (AR) applications, the player or the user's movement often causes the auditory scene to change and the objects to be redistributed relatively. Object-based rendering has already been being used for these applications, with the user movement data playing the role of the rendering control input. Another example in more general audio recording-reproduction chain

is the user's control over one or more individual objects in the total mix, without affecting the others at all. A voice track can be moved, boosted, reduced, or muted completely by the listener, in a song with background music as a mixture of various objects including instruments and vocals, which have not been possible in the channel-based approach without degrading the overall original mix. This enables the user to create his/her own mix, above the producer's intentions, and even above reality if wanted.

Another advantage of the object-based spatial audio could be the wider scalability of the produced material over a variety of rendering systems. The material produced in the object-based manner does not need to consider or to be affected by the rendering system configuration (binaural [11], multichannel loudspeakers, or WFS), since the configuration-specific mixing is carried out by the rendering module. Channel-based production material, however, is fundamentally bound to the rendering configuration. Although it is possible to convert the format at the rendering process (e.g., upmixing/downmixing), the original mix is inevitably altered through this conversion. In the object-based approach, the same material can be used by rendering modules with different configurations for the corresponding output formats, provided that they can interpret the same format of scene description relevantly for their own internal processing—for example, HRTF filtering, Vector-Based Amplitude Panning (VBAP) [10], or WFS processing. Here comes the need for a common scene description format. The Virtual Reality Modeling Language (VRML) standard [6] is known as one of the early forms of such format. Based on this, the Moving Pictures Experts Group (MPEG) later standardized the Binary Format for Scene Description (BIFS) [12] for this purpose. Audio Scene Description Format (ASDF) [4] and Spatial Sound Description Interchange Format (SpatDIF) [9] have also been suggested by other collaborative groups. All of these formats specify the methods to convey the scene description efficiently, using predefined hierarchical entities with features necessary to render the sound objects.

5.4 Challenges of Object-Based Approach

Despite the advancement and potential as the next step of the spatial audio technology, the object-based approach has some challenges to overcome to completely replace the conventional method for spatial audio capturing and production. One of the key challenges is the increasing data size for increasing number of objects to deliver. While in the conventional channel-based production the amount of data is already determined by the number of channels at the rendering side regardless of the number of objects, in the object-based approach the number of objects included in the captured auditory scene determines the amount of data, which can exceed the number of channels in typical multichannel loudspeaker configurations. Additional encoding (compression) would therefore be desirable, particularly in applications requiring streaming, to save the bandwidth. Advanced Audio Coding (AAC) [7], standardized in the MPEG-2 and MPEG-4 specifications, can be used for this,

supporting compression of up to 48 channels of interleaved audio. Spatial Audio Object Coding (SAOC) [5] in the more recent MPEG-D specification is particularly designed for object-based audio storage or transport. In SAOC, individual audio objects are downmixed into mono or stereo signals to minimize the media size and the bandwidth. Meanwhile, side information is prepared which describes the properties of all audio objects and stored as bitstream in addition to the downmixed signals. This side information is used later at the decoding stage for the retrieval of the audio objects. These coding techniques help to reduce the bitrate in the case of a large number of objects to be included in the scene.

Another challenge in the object-based approach is the complexity of the material production. Firstly, if a large number of objects (e.g., instruments in an orchestra) need to be captured simultaneously, obtaining the individual object signals with microphones, without interference from those of adjacent objects, can be difficult, depending on the scene characteristics. A large number of microphones can also be problematic in audiovisual production, if somehow the microphones should not be noticeable. Secondly, if the captured objects are moving, the scene description needs to contain all the trajectories of all captured objects. Authoring of such scene descriptions may not be possible, particularly for broadcasting applications where "live" streaming operation is often required. A solution for the total live operation of the object-based processing chain would need to involve some type of tracking technology, leading to automatic and real-time authoring of the complex scene description. This keeps the object-based approach at the current stage from replacing the channel-based method for the whole broadcasting applications range.

5.5 Summary and Conclusion

The concept of object-based spatial audio capturing to rendering approach has been introduced, with its advantages and challenges as the next step in the spatial audio technology. Its nature of capturing, encoding, and storing or streaming individual sound objects instead of channels as pre-rendered mixture of objects has been described. The meaning of scene description, a feature necessary in the object-based approach for correct final mixing and rendering in various configurations, has been introduced. The user's control over separate objects in the final mix without affecting the others has been described as an advantage that enables interaction between the user and the produced auditory scene. The outlook of using a common scene description language that allows the same production content to be used for multiple rendering systems has also been described as another beneficial point compared to the channel-based approach. Current challenges of this approach, such as larger media sizes for a large number of objects, and the increased complexity in capturing the individual objects separately and in generating the scene description particularly for broadcasting applications, have also been described. However, it has been expected that currently available multichannel or multi-object coding techniques will enable efficient data saving

and that relevant tracking technology along with scene-specific capturing equipment arrangement will enable live production of the scene description and the object-based audio material. With these developments foreseen, the object-based approach is promising as the advanced method of delivering spatial audio for the future, with improvements in general computing power and in data transfer speed.

References

1. Begault DR (1994) 3-D sound for virtual reality and multimedia. AP Professional, Boston
2. Berkhout AJ, DD V, Vogel P (1993) Acoustic control by wave field synthesis. J Acoust Soc Am 93(5):2764–2778
3. Blauert J (1997) Spatial hearing: the psychophysics of human sound localization, 2nd revised edn. Mit Press, Cambridge, MA
4. Geier M, Ahrens J, Spors S (2010) Object-based audio reproduction and the audio scene description format. Organ Sound 15(03):219–227. doi:10.1017/S1355771810000324
5. Herre J, Purnhagen H, Koppens J, Hellmuth O, Engdegård J, Hilper J, Villemoes L, Terentiv L, Falch C, Hölzer A, Valero ML, Resch B, Mundt H, Oh H-O (2012) MPEG spatial audio object coding-the ISO/MPEG standard for efficient coding of interactive audio scenes. J Audio Eng Soc 60(9):655–673
6. ISO/IEC (1997) Information technology—computer graphics and image processing—the virtual reality modeling language (VRML)—part 1: functional specification and UTF-8 encoding. ISO/IEC 14772–1:1997. International Standards Organization (ISO)
7. ISO/IEC (2006) Information technology—generic coding of moving pictures and associated audio information—part 7: advanced audio coding (AAC). ISO/IEC 13818–7:2006(E). International Standard Organization (ISO)
8. Malham D (2008) Spatial hearing mechanisms and sound reproduction. Music Technology Group, University of York. http://www.york.ac.uk/inst/mustech/3d_audio/ambis2.htm. Accessed 17 Mar 2013
9. Peters N, Lossius T, Schacher JC (2012) SpatDIF: principles, specification, and examples. In: The 9th sound and music computing conference, Copenhagen, Denmark. pp 500–505
10. Pulkki V (1997) Virtual sound source positioning using vector base amplitude panning. J Audio Eng Soc 45(6):456–466
11. Rumsey F (2011) Whose head is it anyway? Optimizing binaural audio. J Audio Eng Soc 59(9):672–675
12. Scheirer ED, Väänänen R, Huopaniemi J (1999) AudioBIFS: describing audio scenes with the MPEG-4 multimedia standard. IEEE Trans Multimed 1(3):237–250

Part II
Networking Aspects for 3D Media

Chapter 6
Transport Protocols for 3D Video

Athanasios Kordelas, Ilias Politis, and Erhan Ekmekcioglu

Abstract The multi-view three-dimensional video streaming is emerging as an important technology for future multimedia services. During the last decade, an explosive growth of video applications over the Internet has occurred, resulting to the demand of better network performance. This includes the need to extend video coding and transport protocols as well as to discover new protocols that will guarantee the delivery of real-time high-quality 3D video to consumers. This chapter provides an extended overview of the most widely used media transport protocols MPEG2-TS and RTP on top of scalable video bitstreams. Nevertheless, these protocols have several limitations and are missing key functionalities. Hence, this chapter provides a description of the most recent developments in the MPEG media transport protocol, which is emerging as the next state-of-the-art application layer transport protocol.

6.1 Introduction

MPEG developed the MPEG2 transport stream (TS) [1] in the early 1990s, and since then it has been widely used in media delivery systems such as cable, satellite, and terrestrial delivery of video. Today's digital broadcast channels as specified by the Digital Video Broadcasting Project (DVB) [2] or the Advanced Television Systems Committee (ATSC) [3] rely on MPEG-2 systems for encapsulation and signaling for media delivery. Although over the years new encoding standards such

A. Kordelas (✉) • I. Politis
Department of Electrical & Computer Engineering, University of Patras, Patras 26500, Greece
e-mail: athankord@tesyd.teimes.gr

E. Ekmekcioglu
I-Lab Multimedia Communications Research Group, University of Surrey, Guildford, Surrey GU2 7XH, UK
e-mail: erhan.ekmekcioglu@surrey.ac.uk

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_6, © Springer Science+Business Media New York 2014

as H.263 [4], MPEG-4 visual Part 2 [5], and H.264/MPEG-4 Advanced Video Coding (AVC) [6] and Scalable Video Coding (SVC) [7] have been developed, no significant requirement to update the delivery chain based on MPEG-2 transport stream (TS) has arisen.

On the other hand, the standardization organizations such as IETF, IEEE, and 3GPP have been developing a number of protocols that define the delivery of packetized multimedia content over IP networks. In particular, the real transport protocol (RTP) [8] is an application layer protocol responsible for transporting real-time media data over IP using a variety of underlying transport layer protocols (e.g., TCP [9], UDP [10], and DCCP [11]). The main functionalities provided by RTP include real-time media encapsulation, media synchronization, basic quality of service signaling, and basic coordination for multiparty communication. This makes RTP suitable for different scenarios such as live broadcasting, on-demand multimedia streaming, and other conversational services. Recently, RTP has been updated in order to accommodate forward error correction (FEC) mechanisms and advanced signaling about the received media quality as well as security [12].

The evolution of both standards over the past years has been fuelled by the convergence of broadcasting and mobile services as well as the integration of heterogeneous networking access technologies. The result of this new era in networking is characterized by the need for high availability of both networks and services, in addition to seamless multimedia service continuity across them. Recently, MPEG has been working towards defining a more efficient media transport standard called MPEG media transport (MMT) that aims at addressing all the inefficiencies and disadvantages of current media transport protocols [13].

The rest of the chapter discusses the existing multimedia transport standards MPEG2-TS and RTP with the emphasis on scalable 3D video delivery and includes a presentation of the functionalities and mechanisms described in the current MMT working draft.

6.2 Existing Media Delivery Standards

6.2.1 MPEG-2 Transport Stream

6.2.1.1 General Principle

The MPEG-2 transport stream format specification (MPEG-2 TS, or TS) describes how to combine one or more elementary streams of video and audio, as well as other data, into single or multiple streams that are suitable for storage or transmission. MPEG transport stream is widely used and is the basis for digital television services and applications [14].

A transport stream can be a combination of several programs (e.g., TV channels), where each program comprises elementary streams that share a common time base. A transport stream is a continuous succession of fixed size and mixed

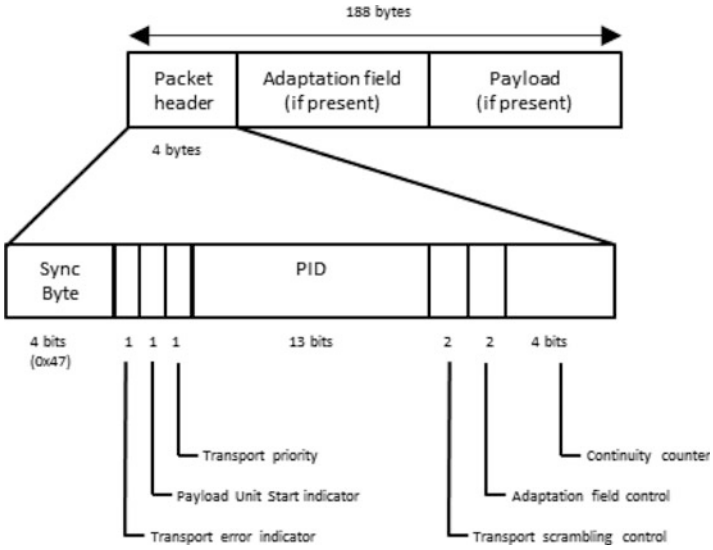


Fig. 6.1 MPEG2-TS packet header

(e.g., audio, video, or related data) packets. Several types of elementary streams can be encapsulated, such as MPEG-2 video (ISO/IEC 13818-2), H.264/ MPEG-4 Part 10 AVC (ISO/IEC 14496-10), SVC and MVC, MPEG-2 Part 7 Advanced Audio Coding (ISO/IEC 13818-7), as well as other information such as metadata and electronic program guides (EPG), including proprietary information. Each such elementary stream type has been defined in MPEG-2 TS standard, and the addition of a new supported format requires an amendment to ISO/IEC 13818-1. An example is the new HEVC standard [15].

A transport stream packet starts with a sync byte and a header as depicted in Fig. 6.1. Additional optional transport fields, as signaled in the optional adaptation field, may follow. The rest of the packet consists of payload. TS packets are 188 bytes in length among which 4 bytes are dedicated to the header. For recorded media applications (such as Blu-ray disks), an optional 4 bytes header can take place before the sync byte, carrying time stamp information.

The first byte of the TS packet header is the synchronization byte (0x47 in hexadecimal representation). Transport error indicator bit determines if the TS packets contain errors after the error correction process (e.g., Reed-Solomon FEC in DVB-S). Payload unit start indicator signals that the first byte of the TS payload is also the first byte of a packetized elementary stream (PES), which includes an elementary stream. The continuity counter is incremented along successive TS packets belonging to the same packetized elementary stream and is used to detect packet losses.

A 13-bit packet ID (PID) identifies each table or elementary stream in a transport stream. A demultiplexer extracts elementary streams from the transport stream in part by looking for packets identified by the same PID. Apart from 17 reserved

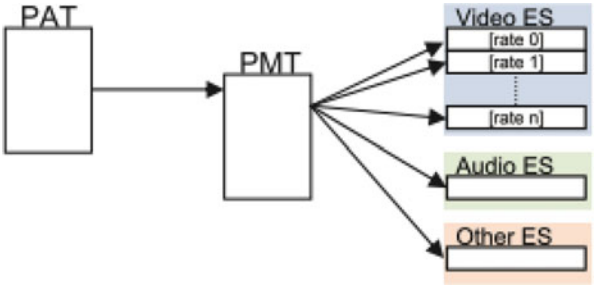


Fig. 6.2 PAT and PMT relationship in MPEG2-TS

values for specific purposes, the PID can take one of the 8,175 available values, which corresponds to the maximum number of elementary streams that can be present in a transport stream.

Within the 17 reserved values, a TS packet exists with a unique PID value of 0x0000 called Program Association Table (PAT) that lists the PIDs of the tables called Program Map Table (PMT) describing each program.

Each single program is described by a PMT in which the PIDs of the TS packets associated with that program are listed. PAT and PMT relationship in MPEG2-TS is depicted in Fig. 6.2.

When the data transmission scheme imposes strict constant bitrate requirements on the transport stream, some additional packets containing no useful information (i.e., stuffing) can be inserted. Usually, PID 0x1FFF is reserved for this purpose.

To enable a decoder presenting synchronized audio and video content, a program clock reference (PCR) is transmitted in the adaptation field of an MPEG-2 transport stream packet. The PCR is associated to a unique PID and is identified by the pcr_pid value in the corresponding Program Map Table (PMT).

The transport stream system layer is divided into two sub-layers, one for multiplex-wide operations (the transport stream packet layer) and another for stream-specific operations (the PES packet layer). Packetized elementary stream is a logical construction and is not defined as a stream for interchange and interoperability purposes. For transport streams, the packet length is fixed, while both fixed and variable packet lengths are allowed for PES, which may be larger than TS packet length. A PES packet consists of a header and a payload, which are data bytes of the elementary stream. There are no requirements to align the access units (i.e., the video frames in the case of a video elementary stream) to the start of a PES packet payload. A new access unit can start at any point in the payload of a PES packet, and several access units can be contained in a single PES packet. A PES packet header starts with a prefix code followed by the stream_id that allows distinguishing PES packets that belong to the specific elementary stream types within the same program. For SVC video streams, all video sub-bitstreams of the same video stream shall have the same stream_id value. Following bits include the PES packet length and two flags indicating the presence of optional information such as copyright information or times stamps (PTS/DTS). Time stamps are used to ensure correct synchronization between several elementary streams of the same program.

The value of the program clock reference (PCR) is employed to generate a system time clock (STC) in the decoder. The STC provides a highly accurate time base that is used to synchronize audio and video elementary streams. This synchronization uses the PTS/DTS (presentation time stamp/decoding time stamp) information. Clocks used at the multiplexer and decoder are measured in units of 27 MHz coded in 42 bits. Each program of a multiplex has its own independent clock which needs not to be synchronized with clocks of other programs of the multiplex. PCR must appear in the multiplex at least every 0.1 s. The presentation time stamp specifies the time in which the decoded access unit has to be presented to the viewer. The decoding time stamp (DTS) specifies the time an access unit has to be decoded. DTS is necessary when B pictures are present in the video elementary stream, since their order of arrival at the decoder is not the same as the presentation order. DTS is always accompanied by PTS. PTS needs to be equal to or greater than its associated DTS. Usually, PTS can be used alone in the elementary stream. Time stamps are expressed in units of 90 kHz and coded in 33 bits.

6.3 Multiplex of MPEG-4 Scalable Video Coding Elementary Streams

SVC is part of MPEG-4 AVC/H.264 specification and detailed in Annex G of ISO/IEC 14496-10 | ITU-T H.264. SVC provides scalability on top of an AVC bitstream that is used as a base layer and can be decoded independently from the enhancement layer(s). Thus, an SVC bitstream is backwards compatible with AVC.

SVC video data is included into NAL (Network Abstraction Layer) units (NALUs) called Video Coding Layer (VCL) NAL, each of them being a packet containing an integer number of bytes. The first byte of each NAL unit is a header byte that contains an indication of the type of data in the NAL unit, and the remaining bytes contain payload data of the type indicated by the header.

NAL units contain video data or other information types, such as sequence parameter set (SPS), picture parameter set (PPS), and optional supplemental enhancement information (SEI). The SPS contains important information that is necessary for the decoding of the video sequence, whereas the PPS contains important information that is necessary for the decoding of one or more pictures in the sequence. SEI has been described in detail in Annex E of ISO/IEC 14496-10. SVC requires the addition of SPS extension and PPS for the extension slices.

A video sequence consisting of a number of successive access units with an SPS can be decoded independently from any other coded video sequence. Each encoded picture belongs to one single access unit that can be signaled by the use of an access unit delimiter NAL (NAL type 9). VCL NAL units of an access unit consist of a set of slices belonging to the same picture. At the beginning of a coded video sequence, the first picture is included in an IDR (instantaneous decoding refresh) access unit. An IDR access unit contains an intra-picture, decodable independently from the

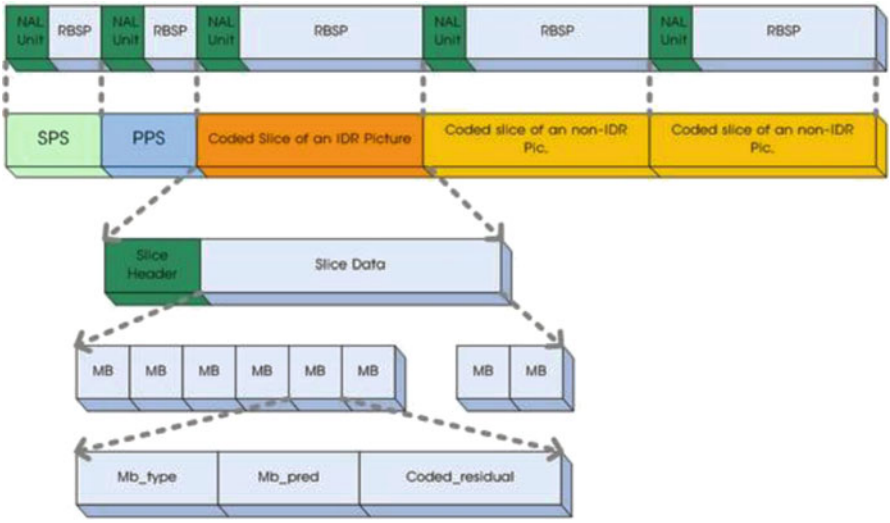


Fig. 6.3 MPEG-4 AVC bitstream structure

other pictures that can be followed by either other IDR access units or non-IDR (N-IDR) access units and contains pictures encoded by using prediction mechanism.

Each NAL unit has one byte or three bytes of header depending on the type of the NAL unit payload of variable byte length, which is called the raw byte sequence payload (RBSP). The overall MPEG-4 AVC bitstream structure is depicted in Fig. 6.3.

SVC NAL units that carry parameter data or RBSP of spatial base layer (H.264/MPEG AVC compatible) need only one byte header to describe the NAL unit type. This NAL header consists of one forbidden bit (F), two bits indicating whether the NAL unit is used for prediction, or not, and five bits to indicate the type, as depicted in Fig. 6.4. At the end of the payload, trailing bits are used to adjust the payload size to become a multiple of bytes.

NAL units of type 20 indicate the inclusion of SVC enhancement layer. They need three extra bytes in their header to describe scalability related information, where:

- R is a reserved bit that if set to 1 signal the presence of SVC NAL unit header.
- I is idr_flag, equal to 1 specifies that the current coded picture is an IDR picture otherwise equal to 0.
- N is no_inter_layer_pred_flag specifies whether interlayer prediction may be used for decoding the coded slice.
- Priority_id specifies a priority identifier for the NAL unit.
- Dependency_id specifies a dependency identifier for the NAL unit.
- Quality_id specifies a quality identifier for the NAL unit.
- Temporal_id specifies a temporal identifier for the NAL unit.

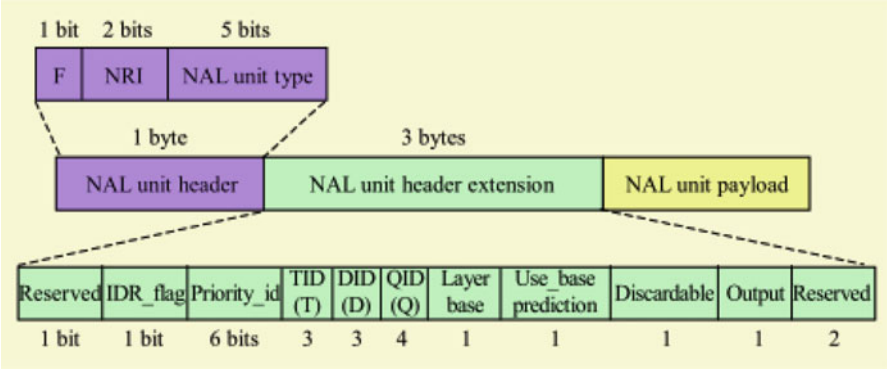


Fig. 6.4 SVC NALU structure

- U is use_ref_base_pic_flag, equal to 1 specifies that reference base pictures (when present) and decoded pictures (when reference base pictures are not present) are used as reference pictures for inter prediction.
- D is discardable_flag equal to 1 specifies that the current NAL unit is not used in the decoding process of NAL units of the current coded picture and all subsequent coded pictures that have a greater value of dependency_id than the current NAL unit
- O is the output_flag 1bits affects the decoded picture output process.
- RR are reserved 2bits, shall be equal to 3.

The SVC base layer, i.e., the AVC video sub-bitstream, has a dependency_id equal to 0. Aspect ratio, timing information, picture colorimetry and picture chrominance locations, and picture structure information of an SVC system are signaled by using video usability information (VUI) parameters inside the sequence parameter set. Even if SEI is optional to decode the video information, it provides useful information, which can be mandatory for some applications like Blu-ray and 3D on smartphones that uses side-by-side representation. All SEI messages that apply to SVC enhancement layers should be included in the AVC video base layer.

The use of MPEG-2 TS as a transport layer for AVC/SVC requires NALUs as an ordered stream of bits, within which the locations of NAL unit boundaries need to be identifiable by a pattern to prevent emulation of the corresponding pattern within the compressed data. AVC/H.264 specification defines a byte stream format (Annex B of ISO/IEC 14496-10 | Rec ITU-T H.264) in which each NAL unit is prefixed by a synchronization byte sequence (0×000001 or 0×00000001). This byte sequence is called a start code prefix and is inserted before every NAL unit to create a new MPEG-4 AVC/ H.264 elementary stream ready to be multiplexed into a MPEG-2 transport stream.

The boundaries of the NAL unit are identified by the unique start code prefix pattern. The most basic NAL unit types related to video streams are listed in Table 6.1.

Table 6.1 ROMEO related NAL unit types

Content of NAL unit	nal_unit_type
Coded slice of a non-IDR picture	1
Coded slice of an IDR picture	5
Supplemental enhancement information (SEI)	6
Sequence parameter set (SPS)	7
Picture parameter set (PPS)	8
Access unit delimiter	9
Sequence parameter set extension	13

6.3.1 Real-Time Transport Protocol

Real-time transport protocol (RTP) provides both unicast and multicast end-to-end network transport functions suitable for applications transmitting real-time data over network services, such as audio, video, or simulation data. The IETF specification on the RTP payload format inherits the basic characteristics of MPEG-4/H.264 AVC in order to enable the packetization and transmission of scalable video formats, such as SVC and Multiview Video Coding (MVC) [16]. Applications typically run RTP on top of UDP to make use of its multiplexing and checksum services, while the RTP header enables the receiver to reconstruct and synchronize the sender's packet sequence without guaranteeing quality of service (QoS). RTP specifies three payload structures for the encapsulation of video NAL units into RTP packets in such a way that fits to the maximum transmission unit (MTU). Each payload structure defines a specific packet type, as shown in Table 6.2.

The first payload structure is the single NAL unit (SNU), which enables the encapsulation of one NALU per RTP packet by using only the RTP header. It should be mentioned that all receivers should support the SNU mode in order to provide backwards compatibility. Moreover, the packet type is specified by the encapsulated NALU type.

The aggregation packets allow the encapsulation of multiple NALUs in smaller sizes (e.g., parameter sets, SEI NALUs) into one RTP packet, reducing the total overhead. Each aggregation packet consists of the RTP header and an "aggregation packet" header that has the same form as H.264/AVC NALU header. Finally, multiple aggregation units are encapsulated, where each aggregation unit encapsulates a NALU using either one or two additional headers depending on its version. There are four aggregation unit versions inherited by MPEG-4/H.264 AVC RTP payload format, as well as one introduced in SVC and passed to MVC, are shown in Table 6.2. It must be mentioned that in the single-time aggregation packet (STAP) mode, the aggregated NALUs must have the same NALU times, while in the multi-time aggregation packet (MTAP) mode, the aggregated NALUs may also have different NALU times.

Finally, fragmentation units (FUs) enable the fragmentation of larger NALUs into smaller RTP packets in order to prevent the IP fragmentation. This scheme increases the robustness with the cost of increased overhead. There are two different FU versions called FU-A and FU-B which includes the decoding order number

Table 6.2 RTP packet types

Packet type (Dec)	Packet type name
0	Reserved
1–23	Single NAL unit
24	Single-time aggregation packet-A (STAP-A)
25	Single-time aggregation packet-B (STAP-B)
26	Multi-time aggregation packet 16 (MTAP16)
27	Multi-time aggregation packet 24 (MTAP24)
28	Fragmentation unit-A (FU-A)
29	Fragmentation unit-B (FU-B)
30	Single NAL unit (PASCI NAL unit)
31	Non-interleaved multi-time aggregation packet (NI-MTAP)

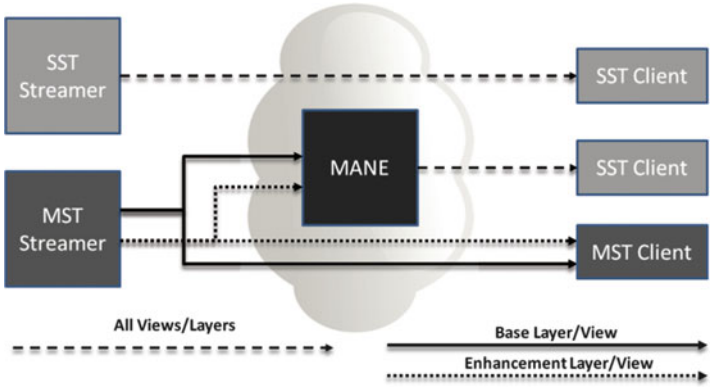


Fig. 6.5 RTP transmission modes for scalable videos

header. In both versions, each FU packet must contain the RTP header as well as two additional headers (one byte long each), the FU indicator and the FU header. From the packetization point of view, the values of the first two fields of the "FU indicator" (F, NRI) are obtained from the corresponding fields of the NALU header, while the last field (type) indicates the packet type (28 or 29). In the same manner, the first two fields of the "FU header" (S, E) indicate the first and the last fragment of a NALU, the third field (R) is a reserved field, and finally the type field obtains its value from the corresponding field of the NALU header. Finally, the H.264/AVC NALU header must be erased since all its values are transferred to the FU headers as mentioned above.

H.264/SVC introduces three transmission modes that are also inherited by H.264/MVC. Figure 6.5 illustrates the three available transmission modes.

In the single-session transmission (SST) mode, all data are carried in a single point-to-point unicast RTP session, using one transport address. Each single RTP session may either carry the base view/layer bitstream or may aggregate more views/layers, depending on the client's needs. This mode is used whenever the potential benefits of MST mode are fewer than the added complexity of the system.

In the case of multi-session transmission (MST) mode, the layered bitstream is transmitted over multiple RTP sessions, each one associated with one RTP stream. Each RTP stream may either carry the base view bitstream or non-base view(s) bitstream(s) or may also aggregate more than one views or layers. MST mode should be used in multicast transmission whenever different receivers must be able to request different number of views of a scalable bitstream.

Finally, a Media-Aware Network Element (MANE) [17] based transmission system can be considered as a middle-box, which is capable of dropping the less important packets of a scalable stream. In this transmission mode, streamer continues to use the MST transmission, but the RTP packets are collected and depacketized from the MANE. With the use of an adaptation decision-taking engine (ADTE), RTP packets may be dropped by taking into account the network conditions and the client's characteristics (i.e., available bandwidth, bit error rate). Finally, the NALUs are re-packetized and transmitted through a single RTP session to the clients. The need for such an intermediate system is stemmed from the existing limitations (e.g., firewalls, NAT protocols) present in real network environments.

During the depacketization process, a client can use the packet type field in order to recognize the payload structure of the packets. The packet type always lies at the first octet after the RTP header (type field), which is always structured as a H.264/AVC NALU header. Generally, the depacketization procedure follows the reverse procedure, but there are a few points that each client should handle:

- Packet reordering is executed using the packet sequence numbers (losses or out-of-order packet arrival can be recognized).
- New frames are recognized by the RTP time stamps.
- In MST mode, the synchronization of the frames received via different RTP sessions is achieved by an additional synchronization process which uses the RTP and NALU headers (i.e., RTP time stamp, nal_unit_type, VID, TID, PRID).

6.4 MPEG Media Transport

6.4.1 Overview

In order to overcome the key functionalities that are missing from both RTP and MPEG2-TS, ISO/MPEG is developing the MPEG media transport (MMT) as the next-generation media transport standard. The scope of MMT working draft is to provide transportation of MPEG media for emerging service over IP networks. Towards this end, the ongoing standardization process has already identified a number of open issues that the new standard aims to address.

In particular, MMT working draft is required to provide efficient mechanisms for delivering emerging applications and contents including 3D video, ultrahigh-definition content, and multi-device and interactive services, adaptively, over

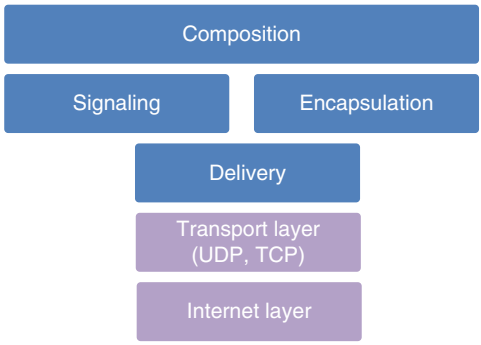


Fig. 6.6 MMT protocol stack

all-IP and broadcast (terrestrial, satellite, and cable) networks. In line with the current tendency for converged networks and services, MMT is designed to seamlessly utilize heterogeneous networks, peer-to-peer networks, and multichannel delivery. Moreover, MMT aims to inherently provide guarantees for QoS and quality of experience (QoE) in media delivery and consumption, by incorporating a cross-layer design. Finally, MMT addresses the demand for transparency to multiple content protection and rights management through an efficient signaling functionality.

The MMT specifies four different functional areas, composition, encapsulation, delivery, and signaling, as shown in Fig. 6.6, that enable the multimedia delivery services. The composition function defines the presentation of MMT-encapsulated content; the encapsulation function defines the format for the encapsulation of encoded media data in order to be either stored or carried as payload of delivery protocols and networks. The delivery function specifies the formats and mechanisms for the transfer of encapsulated media data from one network to another. The signaling function provides signal and control functionality for the delivery and consumption of the media.

Moreover, MMT working draft defines the logical structure of the content. In particular the following terms are defined:

- Media fragment unit (MFU)—a generic container format independent of any specific codec. It consists of media fragment data and additional information that can be used by the underlying delivery layers to control delivery. An MFU can be either a slice or a picture for video.
- Media processing unit (MPU)—a generic container format independent of any specific code that carries one or more access unities (AU). An MPU is composed by one or more MFUs and can contain either timed data or non-timed data. The process of an MPU by MMT entities includes encapsulation/decapsulation and depacketization.
- Asset—a logical data entity that contains coded media data, which may be grouped as one or more MPUs. Individual Assets that can be consumed by the entity directly connected to the receiving MMT entity can be the MPEG2-TS,

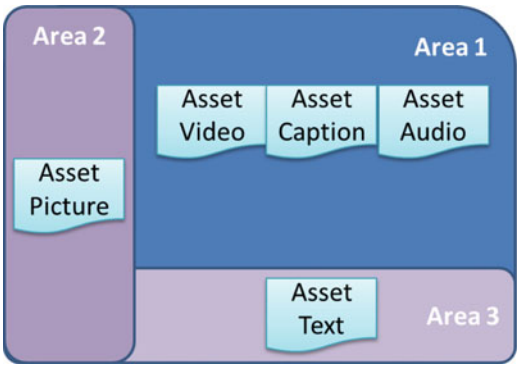


Fig. 6.7 Example of a view as defined by MMT composition structure

PES, JPEG file, etc. An Asset is the data unit for which composition information and transport characteristics are defined.

- Package—is composed by one or more Assets along with additional composition information and transport characteristics. The composition information specifies the spatial and temporal relationship among the Assets and may also determine the delivery order of Assets. The transport characteristics provide the QoS information for transmission and can also be used by the packetizing entity to configure the parameters of the MMT payload and MMT protocol.

6.4.2 MMT Functions

6.4.2.1 Composition

The composition function defines the spatial and temporal relationship among Assets in the Package and includes information that is utilized for delivery optimization and multiscreen presentation of Packages, based on HTML5 with extensions. These extensions to HTML5 include association of Assets in Packages as resource, temporal information that determines the delivery and consumption order of Assets, and, in case of multiscreen environment, mapping of the Assets to a particular screen.

The composition functional area of MMT defines also as view the entire display region that is composed of areas. With the term area, MMT defines a part of a view that is composed of Assets. Figure 6.7 illustrates an example of the view, area, and Asset structural relationship.

6.4.2.2 Encapsulation

The encapsulation function in MMT utilizes the MFU, MPU, and Assets to perform a number of operations similar to PES encapsulation in MPEG2-TS. These

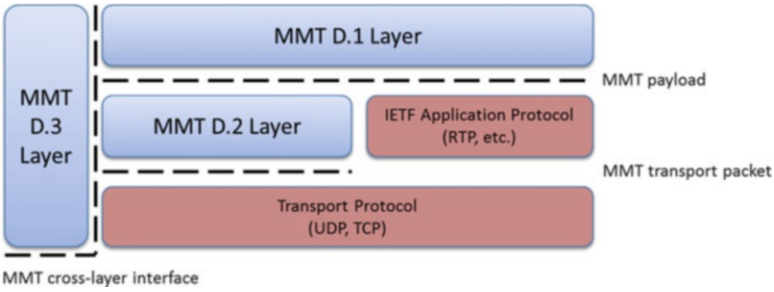


Fig. 6.8 Envision of delivery functional area architecture

operations include media packetization, fragmentation, and multiplexing. Furthermore, encapsulation function is responsible for media synchronization and for providing the composition information that determines the location of media objects in a scene. Additionally, encapsulation defines the parameters for content protection that includes digital rights management and conditional access. The output of the encapsulation is an MMT package.

6.4.2.3 Delivery

The encapsulated MMT package is the input to the delivery functional area that performs the following operations:

- Network layer packetization
- Flow control and multiplexing
- Timing constraints handling and insertion of delivery time stamps
- QoS operations
- Interfaces with application transport protocols (e.g., RTP, RSVP) and transport protocols (e.g., UDP, TCP)
- Error control that includes application layer forward error correction (AL-FEC) and retransmission-based error handling (i.e., ARQ—automatic repeat request for retransmission)

As shown in Fig. 6.8, the delivery function consists of three layers, and each delivery operation is performed in one of these layers.

In particular D.1 layer generates the MMT payload format by indicating a number of functions. The payload identification determines the type of the payload and whether it is media or signaling payload. The fragmentation and aggregation of transport packets information allows the manipulation of the encapsulated media in order to fit the transport layer packets. Finally, D.1 layer information includes the content protection and AL-FEC.

Additionally, the D.2 layer generates the MMT transport packet by inserting in the header information regarding the delivery time stamps and QoS parameters. The MMT working draft specifies the timing model to be used for MMT packets

Fig. 6.9 Signaling functions

delivery, which provides time stamps for synchronization of media streams and calculates network jitter and the amount of delay introduced by the transport and network layers during the MMT package delivery. The protocol supports the delivery of timed media data according to their temporal requirements, as well as, the preservation of the timing relationships among packets in a single MMT Packet flow or among packets from different flows. Currently, MMT protocol assumes that the sending and the receiving entities have access to the universal time clock (UTC) from a remote clock source that utilizes the network time protocol (NTP). However, there are alternative approaches similar to the program clock reference of the MPEG2-TS, which indicate that the delivery clock to be sent in-bound with the media data. Moreover, D.2 layer header includes QoS fields that may be used for network filtering and will eventually be mapped to the corresponding fields of the IPv4 and IPv6 protocols.

The cross-layer optimization in MMT is supported by D.3 layer. The cross-layer function requires exchange of QoS related information between the application and underlying network layers, in order to enable operations such as QoS management and adaptation, flow control, session management, monitoring and error control.

6.4.2.4 Signaling

The signaling function in MMT is divided between the presentation session and the delivery session management layers, as in Fig. 6.9. The first layer defines the formats of control messages exchanged between applications in the client device for media presentation session management and provisioning of required information for individual user consumption. The second layer defines the formats of control messages exchanged between delivery end points and manages the delivery sessions. These signaling messages are used for flow control, delivery session management and monitoring, error control, and hybrid network synchronization control.

6.4.2.5 Error Control

Effective media delivery services over error-prone networks require the implementation of error control, unless TCP is selected as the underlying transport protocol. In the latter case, the packet loss detection and retransmission mechanisms of

TCP are responsible for recovering the lost information. Contrary to using TCP, when error control is not built inside the transport protocol, then any packet loss that occurs during the media delivery is detected at the client device. The MMT working draft defines the application layer FEC mechanism for providing reliable delivery over error-prone IP networks.

In particular, the FEC, which is considered part of the MMT delivery functionality, allows for multi-level construction of MMT packets for both layered and non-layered media data. Hence, this scheme can ensure different levels of protection to each Asset in a FEC source flow. Specifically, two FEC schemes are proposed in the MMT working draft, the first in a two-stage FEC and the second is a layer-aware FEC.

Two-Stage FEC

The two-stage FEC coding structure for AL-FEC is specified to protect a source packet block that contains a predetermined number of MMT packets, as shown in Fig. 6.10.

According to this scheme, every source packet block will be encoded in one of the following FEC coding structures:

- No FEC—no repair blocks are generated.
- FEC 1—one stage FEC, where a single source packet block (i.e., M=1) will be encoded by one of FEC 1 or FEC 2.
- FEC 2—two-stage FEC, where one source packet is split into multiple blocks (i.e., M>1) and the split source packet blocks are converted to source symbol blocks that in terms are encoded by FEC 1 and finally M source symbol blocks are grouped together into a single source symbol block, which is encoded by FEC 2. In parallel, FEC 1 generates M repair symbol blocks R1 (i.e., one for every source symbol block), and FEC 2 generates one repair symbol block R2.

Layer-Aware FEC

The layer-aware FEC (LA-FEC) is proposed for application on layered media (i.e., SVC and MVC) in order to generate as many repair flows as the number of the media layers. LA-FEC exploits the dependency between the base and the enhancement layers and generates repair flows that protect the data of their corresponding layer as well as the data of the layers that this layer depends on. Figure 6.11 illustrates an example of LA-FEC implemented for a base and an enhancement layer.

According to the proposed LA-FEC structure, the MMT packets of the enhancement layers are grouped into source symbol blocks independent of the base layer. The repair flow is then a combination of the source symbol blocks of the corresponding layer and the source symbol blocks from all the layers that the layer depends on.

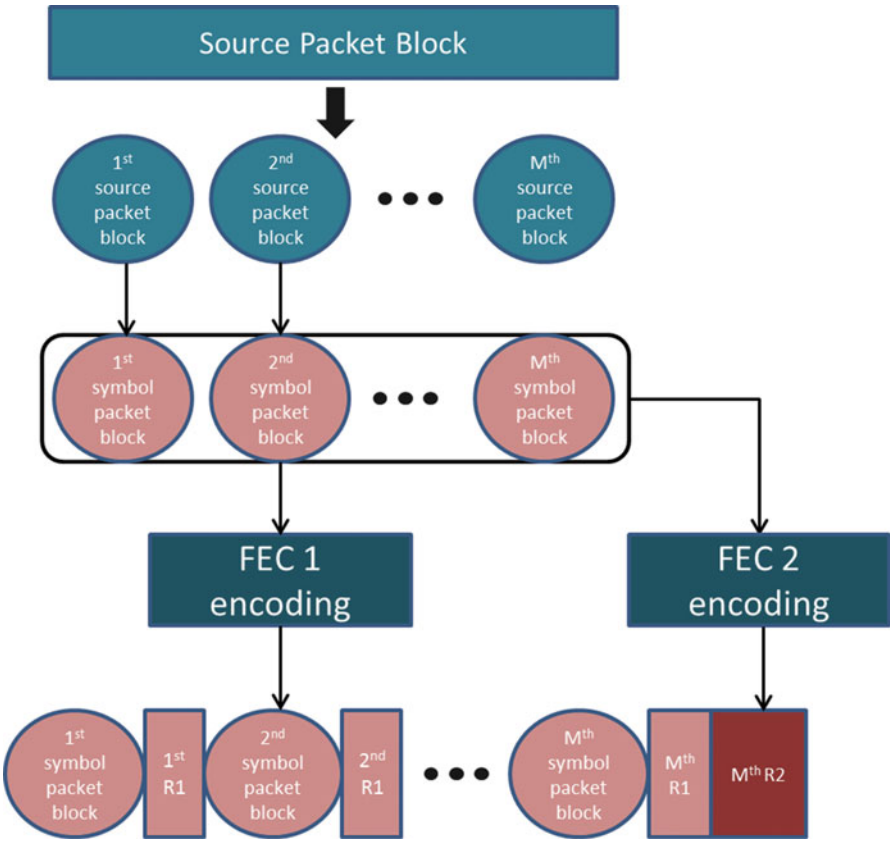


Fig. 6.10 Two-stage FEC structure

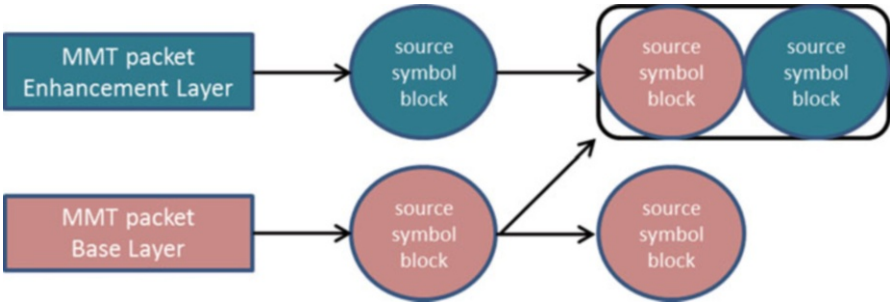


Fig. 6.11 Layer-aware FEC structure

6.5 Conclusions

This chapter has outlined in detail the most widely used media transport protocols, MPEG2-TS, and RTP, with emphasis on scalable video delivery. Although both standards have been updated several times over the recent years, they still have several limitations in terms of quality of service (QoS) guarantee, support for ultrahigh-definition video (UHD), and error control. Additionally, the multimedia user demand for seamless multimedia service continuity over heterogeneous networks renders the current media transport solutions inefficient. Hence, the MPEG media transport (MMT) has been proposed by MPEG as a new application layer transport protocol that efficiently addresses the challenges imposed by immersive multimedia applications and the demanding new networking environment.

References

1. ITU-T Rec. H.222.0 (2007) Information technology—generic coding of moving pictures and associated audio information: systems. May 2006; ISO/IEC 13818-1:2007. Information technology—generic coding of moving pictures and associated audio information (MPEG-2)—Part 1: systems
2. ETSI, E. N. 301 192 (2004) Digital Video Broadcasting (DVB). DVB specification for data broadcasting
3. Dai L, Wang Z, Yang Z (2012) Next-generation digital television terrestrial broadcasting systems: key technologies and research trends. IEEE Comm Mag 50(6):150–158
4. Rijkse K (1996) H. 263: video coding for low-bit-rate communication. IEEE Comm Mag 34(12):42–45
5. Li W (2001) Overview of fine granularity scalability in MPEG-4 video standard. IEEE Trans Circ Syst Video Tech 11(3):301–317
6. Hewage C (2009) Scalable video coding. In: Kondoz AM (ed) Visual media coding and transmission. Wiley, Hoboken, NJ, pp 39–104
7. Schwarz H, Marpe D, Wiegand T (2006) Overview of the scalable H. 264/MPEG4-AVC extension. In: IEEE international conference on image processing, 2006
8. Schulzrinne H (1996) RTP: a transport protocol for real-time applications
9. Postel J (1981) Transmission control protocol
10. Postel J (1980) User datagram protocol. Information Sciences Institute
11. Kohler E et al (2006) Datagram congestion control protocol (DCCP)
12. Watson M, Begen A, Roca V (2010) Forward error correction (FEC) framework. Work in Progress
13. Park K, Lim Y, Aoki S, Fernando G, Lee JY (2012) Information technology-high efficiency coding and media delivery in heterogeneous environments—Part 1: MPEG media transport (MMT). ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio, ISO/IEC CD 23008-1
14. ISO/IEC 13818-1:2007 International Standard (2007) Information technology—Generic coding of moving pictures and associated audio information: systems. [Online] http://www.iso.org/iso/catalogue_detail?csnumber=44169, October 2007
15. Han GJ et al (2012) Overview of the high efficiency video coding (HEVC) standard. pp. 1–1
16. Vetro A, Wiegand T, Sullivan GJ (2011) Overview of the stereo and multiview video coding extensions of the H. 264/MPEG-4 AVC standard. Proc IEEE 99(4):626–642
17. IETF Draft (2011) RTP Payload Format for MVC Video. draft-ietf-payload–rtp-mvc-01. Sept 2011

Chapter 7
Media-Aware Networks in Future Internet Media

Georgios Gardikis, Evangelos Pallis, and Michael Grafl

Abstract Within the Future Internet scene, where multimedia traffic is expected to play a dominant role, the network infrastructure needs to be transformed from a service-agnostic to a media-aware domain, able to offer service-specific handling and in-network operations to the traversing media flows. In this context, this chapter discusses techniques for Media-Aware Networking and illustrates the approaches which have been adopted in the FP7 ALICANTE project towards a novel Media-Aware Network Element (MANE). The media-centric functions of the MANE are presented, namely, content awareness, caching/buffering, QoS management, and flow processing/stream thinning. The added value of media-aware networking is also discussed for four use cases (unicast VoD, multicast streaming, peer-to-peer streaming, and dynamic adaptive streaming over HTTP).

7.1 Introduction

Multimedia (especially video) services constitute a dominant and ever increasing portion of the global Internet traffic, while they are expected to also play a major role in the Future Internet scene [1]. In order to address this reality in the networking domain, a promising perspective is to gradually shift from the current,

G. Gardikis (✉)
NCSR "Demokritos", Institute of Informatics and Telecommunications,
15310 Aghia Paraskevi, Attiki, Greece
e-mail: gardikis@iit.demokritos.gr

E. Pallis
TEI of Crete, Department of Applied Informatics and Multimedia, 71004 Heraklion,
Crete, Greece
e-mail: pallis@pasiphae.teiher.gr

M. Grafl
Alpen-Adria-Universitaet Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria
e-mail: michael.grafl@itec.uni-klu.ac.at

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_7, © Springer Science+Business Media New York 2014

service-unaware, best-effort nature of IP networks into a network logic which is service-aware—and, in specific, media-aware.

This chapter discusses how media awareness can be introduced in the networking domain in a way which is both feasible and scalable, leveraging at the same time state-of-the-art technologies in video representations, such as Scalable Video Coding (SVC) and Dynamic Adaptive Streaming over HTTP (DASH).

7.2 Media Awareness and Media-Centricity

The term *media awareness* refers to the specific handling of each media stream by the network elements, according to its application-layer characteristics, the network conditions, the user demands, and also the various policies/Service Level Agreements (SLAs) which may apply.

Media awareness can be seen as partially correlated with the current trends in Information-Centric Networking (ICN) [2], in the sense that it treats each media flow as a separate information object, with its own characteristics. While the two approaches (media awareness and ICN) do not completely overlap, they could be seen as complementary; media awareness can be exploited in an ICN architecture in order to prioritize, adapt, and cache media content within the forwarding nodes.

Another fundamental difference is that many ICN approaches, although quite promising, assume a clean-slate deployment and thus raise a lot of issues with regard to compatibility with current infrastructures, which directly affect their adoption perspectives. On the contrary, the approach for media awareness [3] presented in this chapter and being designed/implemented in the frame of the EU-funded FP7 ICT ALICANTE project [4] can be deployed in an evolutionary/incremental manner, thus significantly reducing the risks of its adoption and facilitating market penetration. The ALICANTE Media-Aware mechanisms are integrated in an enhanced network node, namely, the Media-Aware Network Element (MANE) [5]. The media-aware functions (Content awareness, Caching, QoS Management and Flow Processing) of the ALICANTE MANE span across OSI Model L3/L4/L7 layers and are summarized in Fig. 7.1.

Content awareness is a broad feature which is primarily based on application recognition, i.e., discrimination between RTP/UDP media streaming, HTTP streaming, and background traffic. This is achieved via heuristic algorithms, able to quickly categorize each type of incoming flows. This basic MANE Flow Classification Framework has been released as open source and can be found in [6].

Going a step further, the MANE gives the Media Service Provider the opportunity to request a more fine-grained classification of the handled flows, without resorting to per-flow signaling mechanisms, which are quite unscalable. This is achieved via a specific signaling mechanism adopted in the ALICANTE project, namely, the Content-Aware Transport Information (CATI) header [7] (shown in Fig. 7.2)

The CATI header consists of 24 bits and is inserted either in the RTP or in HTTP extension headers, depending on the service. (The detailed explanation of each of the fields is beyond the scope of this chapter.) Via the "STYPE" and "SST" fields, the

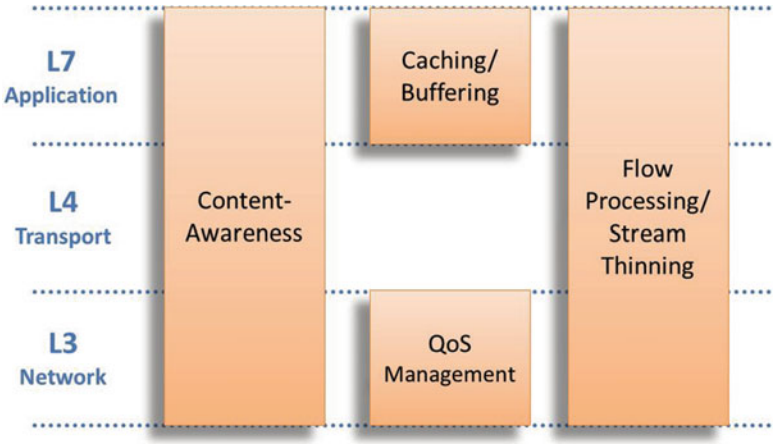


Fig. 7.1 Media-aware functions of the ALICANTE MANE

Fig. 7.2 The content-aware transport information (CATI) header

Media Service Provider can explicitly signal the Service Type and Sub-Service Type to the Network Provider. The MANEs involved in the delivery of the stream parse this header and apply the corresponding policies to the media flow, as have been agreed in the SLA between the Media Service Provider and the Network Operator.

Flow Processing/Stream Thinning (*Media Adaptation*) can be also applied across different layers. Application Layer (L7) adaptation includes full flow processing (media transcoding) and also conversion from single-layer representations to SVC and vice versa (often denoted X-to-SVC and SVC-to-X). Network/Transport layer (L3/L4) adaptation involves dropping and re-including specific SVC layers in a layered stream (stream "thinning," Fig. 7.3). Adaptation decisions depend on the network conditions, the user context and demands as well as CATI (accompanied by the SLA) and are taken by a distributed Adaptation-Decision Taking Framework (ADTF). The exact description of the functionality and logic of the ADTF is beyond the scope of this chapter; more details are to be found in [8].

Caching/buffering mostly applies to P2P and HTTP streaming, where content chunks are cached, according to either the LFU (Least Frequently Used) or LRU (Least Recently Used) policies. This feature gives the MANE CDN-like capabilities, however, focused exclusively on media services. Proactive caching may also be applied, pre-fetching chunks which will be shortly requested, as will be discussed in the Adaptive HTTP case.

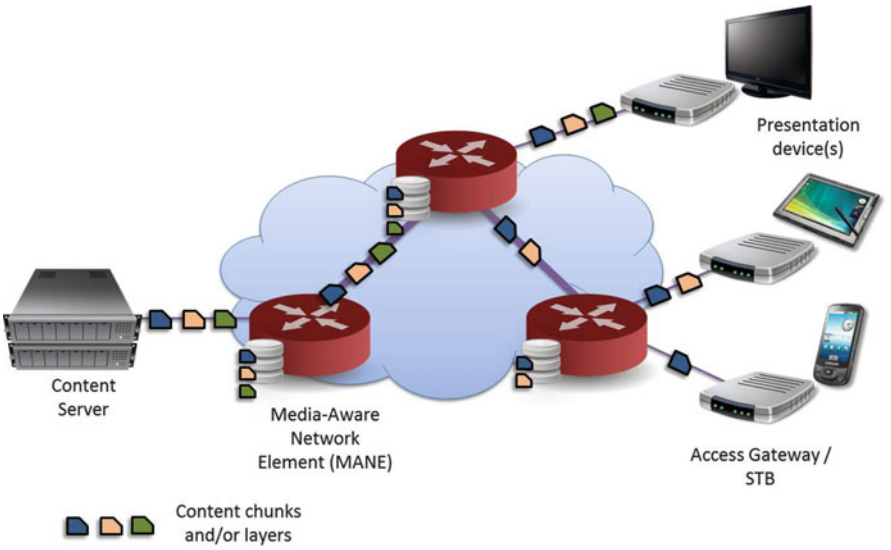


Fig. 7.3 Selective distribution of content layers and/or chunks according to the network conditions, user demand/context, and service-level agreements (conveyed in the CATI)

Last, *QoS Management* refers to the enforcement of QoS policies to each flow, according to the findings of the Content Awareness function and also the network conditions. In the ALICANTE MANE, QoS enforcement is achieved via well-established traffic differentiation mechanisms such as DiffServ.

The following section describes the behavior of the aforementioned MANE functions in four different use cases: unicast Video-on-Demand (VoD), multicast video, peer-to-peer streaming, and dynamic adaptive HTTP Streaming. For a detailed discussion, the interested reader is referred to [8].

7.3 Use Cases for Media-Aware Networking

7.3.1 Use Case 1: Unicast VoD Streaming (RTP/RTSP)

The Unicast VoD use case refers to a single sender streaming the video to a single receiver (also applicable to one-to-one videoconference). We assume the use of RTP for media transport and RTSP for session control; HTTP streaming is discussed in Sect. 7.3.4. In a deployment with SVC, all layers are packed into a single RTP session.

The MANE becomes aware of the properties of the media stream via RTP header inspection and parsing of the CATI header. In case of network congestion, the MANE applies QoS policies to each flow. RTP sessions are always prioritized over

best-effort traffic, even over HTTP streaming. At a more fine-grained level, prioritization among different RTP streams is based on CATI signaling, according to the established SLAs; prioritization can take place among different media service providers and also among different services of the same provider.

For SVC-based transmission, the MANE will furthermore react to network congestion by selectively dropping the higher enhancement layers, always according to the service and flow priority, denoted in the CATI header. This assures continuous playout of at least the base quality at the receiver. Although the content is received at a lower bitrate and the decoded stream is of lower quality, the actual quality of experience is much higher than in the case of a video stalled or full of visual artifacts caused by uncontrolled packet loss. As soon as there is again enough network capacity, the dropped layers can be re-included.

7.3.2 Use Case 2: Multicast Streaming

Multicast streaming involves the streaming of content to multiple receivers simultaneously, using native IP multicast, at least in the core/edge network. In the case of SVC, different multicast trees are constructed for each SVC layer, involving the same root (sender) but different leaves (receivers). The SVC base layer is transmitted to all receivers, while the reception of enhancement layers by each receiver depends on the receiver context (e.g., screen resolution) and also the network conditions. The most straightforward way of achieving this is via receiver-driven layered multicast (RDLM) [9], in which different layers are transmitted over separate multicast groups and over different RTP sessions. Each receiver only subscribes to the layers which it can actually present to the viewer and can also be handled by the access network link.

Similar to the unicast case, the MANE becomes aware of the properties of the multicast stream via RTP header inspection and parsing of the CATI header. In the case of adequate capacity, the MANE can fulfill the aforementioned receiver-driven selective multicast via standard IGMP-based multicast routing; a multicast flow—corresponding to a media stream or a single SVC layer only—is forwarded only if at least one receiver has subscribed to it. Content-aware QoS enforcement is applied only when congestion occurs; in this case, the MANEs will selectively drop the higher enhancement layers, always according to the service and flow priority, as denoted in the CATI header.

7.3.3 Use Case 3: Peer-to-Peer Streaming

In peer-to-peer (P2P) streaming, multiple senders dispatch parts of the media content (called chunks) to multiple receivers. In contrast to traditional P2P file distribution, P2P streaming exhibits timing constraints; every chunk must arrive before its playout deadline expires. Scalable media representation allows the receiver to request only those layers that are supported by the player and by the capacity of

the access link; in this case, peer selection and chunk fetching must be done in a way that the base layer is prioritized and decoded/presented in time [10].

P2P is a use case which leverages MANE in-network media-aware caching. Popular media chunks are locally cached at the MANE and serve multiple client requests; this is especially useful in live-streaming scenarios, where almost all receivers share the same time window for the content; thus, each piece will be highly popular for a short time span. By buffering a piece during this time frame, the MANE is able to reduce network utilization and latency even with a limited buffer size.

7.3.4 Use Case 4: Dynamic Adaptive Streaming Over HTTP

The dynamic adaptive streaming concept, standardized as MPEG-DASH [11], allows media content receivers to self-adapt to the optimal stream bitrate. To achieve this, the content is stored in the server in various resolutions and bitrates (called representations). Each representation is fragmented into segments, which are time-aligned across different representations. A so-called manifest file contains the structure of the media stream, including the description of each representation and the URL of each segment. This mechanism allows the receiver to seamlessly switch between representations, changing stream resolution and bitrate on-the-fly. Furthermore, the use of HTTP alleviates most firewalling issues and allows easy in-network caching by standard Web caches and CDNs. HTTP streaming is typically used in unicast mode, but MANE-assisted multicast or even P2P streaming modes are possible as well.

DASH streams are recognized by the MANE by parsing the HTTP header and especially the CATI field within it. This allows DASH streams to be prioritized over other HTTP (Web) traffic. Prioritization among DASH services is decided according to the CATI field and conforming to the established SLAs. In any case, since HTTP streaming uses TCP for transport, it is more resilient to errors and packet losses than RTP/UDP streams. Thus, the latter are generally assigned a higher priority by the MANE.

In-network adaptation does not apply to DASH streams, since the adaptation procedure is entirely receiver-driven. However, in-network caching at the MANE is greatly promoted by the use of DASH. HTTP streaming can immediately benefit from existing HTTP caching technologies [12], which can be used "as-is," adding CDN functionalities to the MANE. The added value of media awareness relies on proactive caching; having the MANE parsing the manifest file enables it to know a priori which segments will be shortly request it. These segments can be pre-fetched and cached locally for instant delivery. This is especially useful for high-popularity live streams, since it can considerably reduce the viewing latency.

7.4 Conclusions

This chapter presented the added-value features of the MANE, as designed and developed in the frame of the EU FP7 ICT project ALICANTE. The investigation of four separate use cases showed that media awareness at the network domain can introduce significant benefits in the media provision chain by allowing content awareness, service differentiation, as well as in-network media caching and adaptation.

Further evolutions of MANEs would target to a more active participation of the MANE in media delivery by allowing them to evolve to multimodal proxies, combining/bridging different delivery modes. For example, a MANE could be able to request a stream over p2p and then multicast it or translate between RTP and adaptive HTTP streaming.

Acknowledgement This work was supported in part by the EC in the context of the ALICANTE project (FP7-ICT-248652).

References

1. Pan J, Paul S, Jain R (2011) A survey of the research on future internet architectures. IEEE Commun Mag 49(7):26–36
2. Choi J, Han J, Cho E, Kwon T, Choi Y (2011) A survey on content-oriented networking for efficient content delivery. IEEE Commun Mag 49(3):121–127
3. Koumaras H, Negru D, Borcoci E, Koumaras V, Troulos C, Lapid Y, Pallis E, Sidibe M, Pinto A, Gardikis G, Xilouris G, Timmerer C (2011) Media ecosystems: a novel approach for content-awareness in future networks. In: Domingue J et al (eds) The future internet, future internet assembly 2011: achievements and technological promises, vol 6656, Lecture notes in computer science. Springer, Heidelberg, p 369. doi:http://dx.doi.org/10.1007/978-3-642-20898-0_26
4. ALICANTE Web site. http://ict-alicante.eu/
5. Vorniotakis N, Xilouris G, Gardikis G, Zotos N, Kourtis A, Pallis E (2011) A preliminary implementation of a Content-Aware network node. In: Proc. IEEE Int. Conf on Multimedia and Expo (ICME), Barcelona, Spain, 11–15 July 2011, pp 1–6
6. Media Network Aware Element (MANE) Flow Classification and Configuration Framework. http://www.medianetlab.gr/opensource/
7. Li J, Zhang Y, Leib Z, Rodrigues P, Chen Y, Arnaud J, Négru D, Borcoci E, Bretillon P, Renzi D (eds) SP/CP service environment—intermediate. In: ICT-ALICANTE, Deliverable D5.1.1, September 2011
8. Grafl M, Timmerer C, Hellwagner H, Xilouris G, Gardikis G, Renzi D, Battista S, Borcoci E, Negru D (2012) Scalable media coding enabling content-aware networking. In: IEEE MultiMedia, (Preprint) November 2012
9. McCanne S, Jacobson V, Vetterli M (1996) Receiver-driven layered multicast. In: Conference proceedings on applications, technologies, architectures, and protocols for computer communications, Palo Alto, California, United States, October 1996, pp 117–130

10. Eberhard M, Szkaliczki T, Hellwagner H, Szobonya L, Timmerer C (2011) An evaluation of piece-picking algorithms for layered content in Bittorrent-based peer-to-peer systems. In: Proc. IEEE Int. Conf. on Multimedia and Expo (ICME'11), Barcelona, Spain, July 2011
11. Sodagar I (2011) The MPEG-DASH standard for multimedia streaming over the internet. IEEE Multimed 18(4):62–67
12. Sánchez de la Fuente Y et al (2011) iDASH: improved dynamic adaptive streaming over HTTP using scalable video coding. In: Proceedings of the second annual ACM conference on multimedia systems, New York, NY, USA, February 2011, pp 257–264

Chapter 8
P2P Video Streaming Technologies

Konstantinos Birkos

Abstract Peer-to-peer streaming enables end-users to cooperate and exchange available network resources in order to share video content. Peer-to-peer streaming represents a prominent approach that provides scalability, robustness, and efficient resource utilization. This chapter aims to present the basic aspects of peer-to-peer streaming technologies. Mechanisms for establishing connections between peers as well as mechanisms for exchanging video data are analyzed. Theoretical findings that characterize the performance of peer-to-peer streaming systems are also provided.

8.1 Introduction

Peer-to-peer streaming has enabled massive access to video content either on-demand or real-time by exploiting for Internet users. The great advantage of peer-to-peer streaming is the efficient exploitation of the users resources towards the formation of a dynamic infrastructure. Peer-to-peer streaming systems are inspired by their file-sharing equivalents. Therefore they have inherited their main advantages, namely scalability, robustness, self-treatment, resilience, and self-organization which are supported by central entities in some cases.

According to the peer-to-peer approach, the video content becomes initially available from a set of servers (or a set of end-users in case of user-generated con-tent). The servers act as mediators between the video source and the users (real-time streaming) or as storage entities (on-demand streaming). End-users called peers can connect to the server, retrieve the requested video content, and make it available to other peers by means of a predefined diffusion mechanism which is implemented as a computer software called peer-to-peer client.

K. Birkos (✉)
University of Patras, Patras 26500, Greece
e-mail: kmpirkos@ece.upatras.gr

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_8, © Springer Science+Business Media New York 2014

Two basic issues arise from the peer-to-peer streaming approach. The first issue is selecting the peers to receive video content from and the peers to forward the received video content to. The second issue is scheduling the delivery of the video content between peers.

As in the peer-to-peer file sharing paradigm, the outcome of peer selection is an overlay network which is formed by logical connection between peers. The overlay network can be structured or unstructured.

A structured overlay has a fixed structure which is defined by overlay formation and maintenance mechanisms. The overlay formation mechanism dictates the establishment of logical connections between peers. The overlay maintenance mechanism reconstructs the overlay in case of peer departures. In an unstructured overlay, connections between peers are formed dynamically according to certain peer selection criteria. As a result, the traffic flow between peers is also dynamic and there is need for scheduling video delivery among peers.

Tree-based overlay networks are a special case of structured overlays. A tree-based overlay is formed as an application-layer multicast tree on top of which resides the video server. In tree-based overlays, the video content is pushed from the root towards the leaves of the tree. Mesh-based overlay networks rely on the dynamic formation of logical links between peers. The establishment of logical connections is performed upon mutual agreement of peers. Proper mechanism handles the exchange of video portions among connected peers.

Video streaming over peer-to-peer networks poses many challenges [15]. A first challenge is to cope with the negative impact of different topologies on streaming performance. Both tree-based and mesh-based approaches have limitations that stem from inherent topology characteristics. Another challenge is to cope with the heterogeneity of peers in terms of upload and download bandwidth. As proved in [6], heterogeneity can cause performance degradation in peer-to-peer streaming applications. Peer dynamics also have a negative effect on the service capacity [24]. The unpredictable behavior of peers regarding session initiation and termination leads in topology changes which harm video delivery.

The rest of the chapter is organized as follows: Sect. 8.2 elaborates peer-to-peer streaming over tree-based overlay networks. The main approaches for overlay formation and maintenance are presented. Section 8.3 is about peer-to-peer streaming over mesh-based overlay networks. State of the art in peer selection and chunk scheduling is surveyed. Section 8.4 provides a comparison between tree-based and mesh-based peer-to-peer streaming systems. Finally, Sect. 8.5 concludes the chapter.

8.2 Tree-Based Overlay Networks

Tree-based overlay networks have deterministic data delivery paths where the parent and children of a peer are predetermined and thus have more predictable content delivery characteristics.

8.2.1 Basic Principles

In single-tree systems, whole content is distributed over a single-formed tree. In these systems, leaf nodes of the tree are consumers without uploading any content to the other peers, which may be considered as an unfair job division. Besides, in case of a non-leaf peer churn, children of this peer become totally disconnected from the content delivery structure. On the other hand, use of multiple multicast trees may eliminate the unfairness issue by inserting the leaf nodes of one tree as intermediate peer in other trees. Resilience to the peer churn is also improved in multiple multicast tree systems by delivering the data redundantly over different paths.

In each tree, peers are represented by vertices and logical connections between peers are represented by directional edges. The direction of each edge is also the direction of the data flow from a source peer to a destination peer. In each tree, a peer has a single source peer called *parent* and a set of destination peers called *children*. The peer receives video chunks from its parent peer and forwards the received chunks to its children peers. In this way, video chunks are disseminated to all the peers in the tree.

Each peer participates in all the trees. Consequently, each tree is formed by the same set of peers. However, each peer is placed in a different position in each tree. A peer can be repositioned several times throughout the duration of its session due to peer arrivals and departures that alter the formation of the tree dynamically. Therefore, each peer has a different parent and a different set of children peers in each tree. This structure has an inherent resilience to peer dynamics: when a peer is disconnected from a subset of the available trees, it can still receive chunks via the rest of the trees.

In a specific tree, a peer can be *fertile* or *sterile*. A fertile peer can forward chunks to children peers whereas a sterile peer only receives chunks. In order for the overlay to be feasible, every peer has to contribute using its uploading bandwidth. Therefore, a peer is fertile in some trees and sterile in the others. A fundamental feasibility assumption is that the total number of inbound connections to a peer must be equal to the total number of outbound connections. In other words, if a peer participates in t trees, it has t parents and it must offer at least t slots in total. Therefore, if the peer is fertile in d trees, it must offer t / d slots on average in each fertile tree.

At the roof of all trees there is a server that has the role of the root peer. The server is responsible for the structure of the trees and the data flow towards the peers. The server maintains the structure of the trees, monitors the overlay, and distributes video chunks to the trees. In each tree, a number of peers are directly connected to the server. The server sends the same video chunks to its children peers in each tree. Moreover, different video chunks are sent to each tree in a Round-Robin fashion. For instance, if seq_i is the unique sequence number of chunk i, then this chunk is forwarded to the tree $\text{mod}(\text{seq}_i, t) + 1$. During disconnection periods, peers do not receive the portion of video chunks that are propagated through the disconnected trees. An example of the structure of multiple multicast

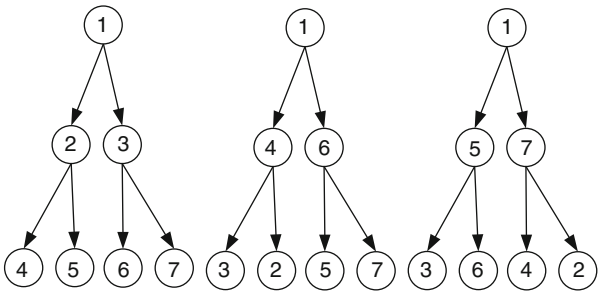


Fig. 8.1 The structure of a multiple multicast tree-based overlay

trees is given in Fig. 8.1. If peer 3 leaves the system, peers 6 and 7 get disconnected from the first tree. However, they are still connected in the other trees.

8.2.2 Overlay Formation and Maintenance

In CoopNet [12, 13], a centralized approach for the overlay control is adopted. The basic characteristics of tree formation and maintenance are:

Peer insertion: When a peer wishes to join the overlay, it contacts the server and requests access. The server finds the d trees with the least fertile peers and assigns the joining peer to be fertile in these trees. The peer is sterile in the remaining $t - d$ trees.

(a) Insertion of fertile peer: the server searches the tree from top to bottom in a breadth-first fashion. The term *level* means the number of logical hops between the server and a peer. The search stops upon finding a fertile peer with free slots or a peer with sterile children. (1) If it finds a peer with free slots in a certain level, it searches the rest of the peers in that level. If there are more peers with free slots, the server connects the joining peer with the fertile peer with the most free slots. Otherwise it connects the joining peer with the first peer found. (2) If the server finds a peer with no free slots but with at least one sterile child, it searches the remaining peers at that level. If peers with free slots are found, the previous action (1) is performed. If no peers with free slots are found and if more fertile peers with sterile children are found, the server connects the joining peer with the peer with the most sterile children. The sterile child is replaced by the joining peer and the sterile child connects with the joining peer. If none of the cases (1) and (2) holds, insertion is impossible and the peer is marked as *disconnected* in that tree. Algorithm 1 shows the algorithm that describes the insertion of a fertile peer in a tree.

Algorithm 1 Insertion of fertile peer

```
repeat
    move to next peer
until a peer with free slots or sterile children is found
  candidate parent = peer
for each peer at this level do
    if (peer has more free slots than candidate parent) or (candidate parent does
    not have free slots and this peer has more sterile children) then
      candidate parent = this peer
    end if
end for
if candidate parent has free slots then
    attach joining peer to candidate parent
else if candidate parent has sterile children then
    detach a sterile peer from candidate parent, attach joining peer to candidate
    parent and attach sterile peer to joining peer
end if
```

(b) Insertion of a sterile peer: the server performs the set of actions in case (1) described previously. It searches only for fertile peers with free slots since the joining peer cannot offer slots in order to replace a sterile peer. If no peers with free slots exist in the tree, the joining peer is marked as *disconnected*. Algorithm 2 describes the insertion of a fertile peer in a tree.

Algorithm 2 Insertion of a sterile peer

```
repeat
    move to next peer
until a peer with free slots is found
  candidate parent = peer
for each peer at this level do
    if peer has more free slots than candidate parent then
      candidate parent = this peer
    end if
end for
attach joining peer to candidate parent
```

Peer deletion: When a peer wishes to leave the overlay, it informs the server. The server considers the peers that depend on the departing peer as *temporarily disconnected* and performs a peer insertion process (a) for each one of them.

Two different policies can determine whether a fertile peer has free slots, namely *balanced* and *unbalanced* allocation. In balanced allocation, a peer has almost the same number of available slots in each tree. In unbalanced allocation, the available

slots of a peer are allocated opportunistically among trees, according to the current needs. In this work, only the balanced policy is considered.

SplitStream [5] is another tree-based overlay. Contrary to CoopNet, in which the server has full control of the overlay formation and maintenance, SpliStream adopts a more distributed approach. In doing so, it relies on Scribe [4] which in turn relies on Pastry [16].

Pastry is a structured overlay network that implements lookup and overlay routing services for retrieving items stored in a distributed hash table (DHT). As in other structured overlays, peers are assigned a unique identifier called *peer ID* and each stored object is assigned a unique key, namely *object key*. Peer IDs and object keys share the same space. When a peer receives a query about an object key, it forwards the query to a peer whose peer ID is numerically closest to the object key. Each peer maintains a routing table that consists of the peer IDs which share some initial bits with the peer ID of the current peer. Therefore, the query is forwarded to a peer from the routing table which has a peer ID that shares at least one bit more with the object key than the current peer.

Scribe exploits the structure of an existing Pastry overlay network in order to provide application-level multicast services. In Scribe, each multicast trees is assigned a unique identifier called *tree ID*. Tree IDs share the same space with peer IDs. Each peer with the same peer ID as a tree ID is root in the corresponding multicast tree. Each multicast tree is formed by the total of Pastry-based routes from each peer to the root peer of the tree.

Pastry is inherently decentralized, i.e. overlay reformulation after peer insertion and peer deletion is not handled by a central entity. Therefore, multicast tree formation and maintenance in Scribe is also decentralized.

SplitStream extends Scribe by adding some extra functionalities. A peer can join any subset of the set of available trees. In order to eliminate correlated paths between different trees, all tree IDs differ in the most significant bit. IDs are chosen from an address space in base 2^b. If the number of trees is equal to 2^b, each peer has equal probability to be fertile in a tree.

The bounded upstream capacity of peers is a constraint that SplitStream addresses as follows: The upstream capacity determines the *maximum peer degree*, i.e. the maximum number of children peers a peer can support. When a peer that has reached its maximum degree receives a join-request message from a new peer, it decides whether it will remove any of its children in order to adopt the new peer or not. Algorithm 3 presents the set of actions performed after a join request is made to a “saturated” peer.

Algorithm 3 Peer with no free slots handling join request from a new peer

```
adopt the new peer as a prospective child
find set of children with minimum match between peer ID prefix and tree ID prefix
if prospective child in set then
    reject prospective child
else
    remove any child in set at random
    accept prospective child
end if
```

If the peer rejects the prospective child, it responds with a list of its current children. The new peer then tries to connect to a peer within this list. This is a recursive process that terminates when the new peer finds an available parent peer. If the orphan peer fails to find a parent, it sends an anycast request to all peers who have available slots. If the are no peers with free slots, the overlay is infeasible. Otherwise, either there is no peer with free slots in the specific tree or the candidate parent is a successor of this peer.

Algorithm 4 Peer trying to connect to a tree with no free slots

```
find a leaf peer in the tree
if the leaf peer has no forwarding capacity then
    request parent of leaf to drop leaf
else if the leaf peer has no free slots in any other tree then
    the leaf peer drops a child at random
    the leaf peer adopts the orphan peer
end if
```

Chunkyspread [20] is another tree-based streaming system. Chunkyspread runs on top of a network formed with the Swaplinks algorithm. Swaplinks produces and continuously refreshes a random graph by means of weighted random walks. Nodes in the random graph represent participating peers. This structure is exploited towards the formation of multiple multicast trees. Neighbor nodes in the random graph are not necessarily connected peers in a tree. Instead, neighbor nodes may possibly have a parent–child relation.

Each peer has a *maximum load* which equals the number of outbound connections. Each peer has also a *target load (TL)* which is the preferred number of outbound connections during steady state. TL must be higher than a predefined threshold called *minimum load (ML)*. Since latency is important in peer-to-peer streaming, Chunkyspread defines *latency threshold (LT)* as the percentage of the target load. Target load reduced by latency threshold percent is called *lower latency threshold (LLT)* whereas target load increased by latency threshold percent is called *upper latency threshold (ULT)*. Each peer must have a minimum number of

outbound connections which is called *minimum node degree (MND)*. MND is actually the minimum number of children in all trees.

The aforementioned parameters are used in the tuning of load balancing and latency. The number of children peers is called *node degree (ND)* and it must be proportional to the target load as follows: $ND = \max[MND, (TL/ML) * MND]$. If a peer's load is below LLT, other peers will possibly try to add the peer as their parent peer. If a peer's load is above ULT, some of its children will try to find other parent peers. Finally, if a peer's load falls within the preferred range, the peer tries to improve latency by changing parents. Dynamic switching between parent neighbors and non-parent neighbors is the main mechanism for both load balancing and latency optimization. Details of these interactions are given next.

When a peer joins the system, it discovers streams traversing different trees via control packets flooded by its neighbor nodes. After that, it selects a parent peer for each tree from the set of neighbor nodes. A peer periodically checks for overloaded parents (with load that exceeds ULT) and underloaded neighbors (with load below LLT). Underloaded neighbors are considered as candidate peers to replace overloaded parents. In this case, the peer informs each overloaded parent of the load of all the candidate parents (underloaded neighbors). This is called a *switch request*. An overloaded parent tries to return to the normal load range and at the same time to preserve load balancing for other peers. In doing so, when it receives a switch request from a child, it checks the provided list of candidate parents–neighbors and it dictates the child to attach to the neighbor with the least load.

If the parents of a peer are not overloaded, the peer checks whether it can switch parents in order to improve latency. A neighbor that receives packets from a tree with lower delay than a parent in the same tree can replace this parent.

8.3 Mesh-Based Overlay Networks

In mesh-based overlay networks there is no constant topology and connections between peers are formed dynamically. Two main problems arise in this case: peer selection and chunk scheduling.

8.3.1 Basic Principles

The participating peers exercise peer selection algorithms in order to select the set of peers to exchange content with. Peer selection strategies define the number of peers to connect with, which peers to select, when and how often to change the selected peers [9]. Usually, when a peer joins the system, a *tracker* provides the peer with a list that contains a random subset of currently active peers. The peer

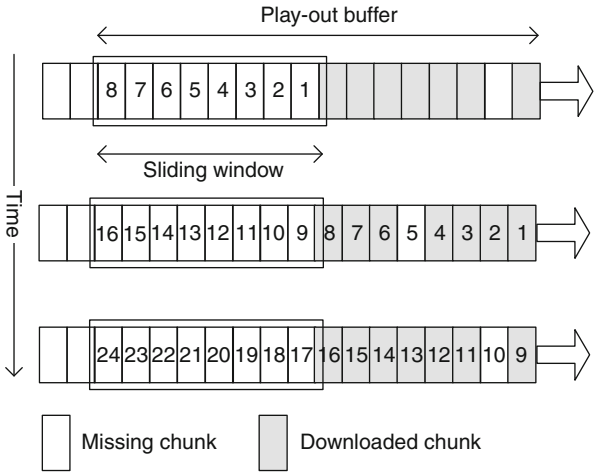


Fig. 8.2 Evolution of available chunks in the chunk buffer

applies the peer selection algorithm in order to select some of these peers as *partners* (or *neighbors*) and start exchange chunks with them.

The exchange of chunks is handled by the chunk scheduling mechanism. The basic idea in the application-layer multicast applied in mesh-based system is that each peer receives missing chunks from neighbor peers that have these chunks stored in their *play-out buffer* and transmits chunks stored in its own play-out buffer to neighbor peers that need these chunks. In the former case, the chunk scheduling mechanism selects which chunks to be delivered to which neighbors, and at which order. In the latter case, the chunk scheduling mechanism selects which chunks to be requested from which neighbors and at which order. The length of the play-out buffer is a system parameter.

There are two main approaches in mesh-based streaming, namely *push-based* and *pull-based*. In systems, a peer forwards the received chunks to neighbors that do not have these chunks. The main drawback in push-based streaming is the redundant chunk pushes: a peer may choose to forward a chunk to a neighbor that already has this chunk. In pull-based systems, redundant pushes are avoided because each peer requests the chunk it needs from its neighbors. This implies that each peer must be aware of the chunks stored in the chunk buffers of its neighbors. This is realized by periodical exchange of *buffer maps* between neighbor peers. A buffer map is a list of the chunks currently stored at a peer's chunk buffer.

Chunk scheduling is performed on a periodical basis. Chunks are divided in *sliding windows* according to their time order. The window length is expressed as number of chunks and it is a system parameter. The chunk schedule (decisions on the which chunks to be requested, from which neighbors and at which order) is defined and executed at the end of each window. Moreover, neighbor peers exchange buffer maps at the beginning of each window. The information of the chunk availability is used as input to the chunk selection algorithm. Figure 8.2

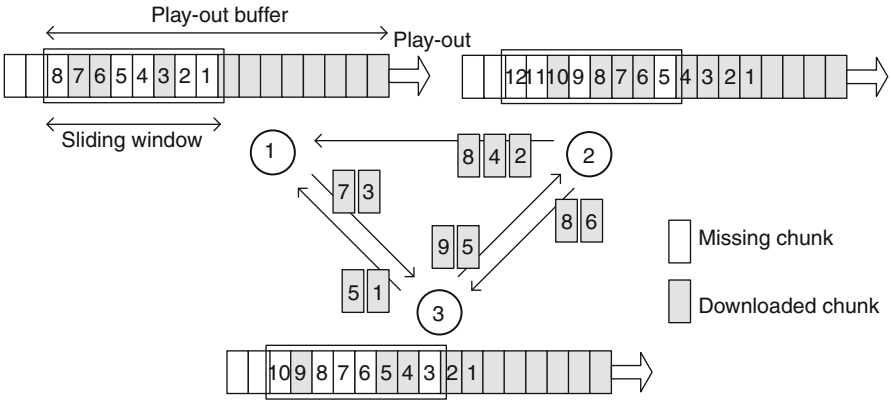


Fig. 8.3 An example of pull-based streaming over mesh-based overlays

presents an example of the evolution of the available chunks in the play-out buffer of a peer. In this example, the size of the play-out buffer is 16 and the size of the window is 8. In each round, a number of stored chunks are dequeued and consumed by the video player. At the same time, requests for new chunks within the sliding window must be scheduled.

Figure 8.3 presents an example of pull-based streaming. Neighbor peers 1,2 and 3 exchange buffer maps and schedule chunk requests between each other. Peer 1 pulls chunks 2,4,8 from Peer 2 and chunks 1,5 from Peer 3. Peer 2 pulls chunks 5,9 from Peer 3. Finally, Peer 3 pulls chunks 3,7 from Peer 1 and chunks 6,8 from Peer 2.

8.3.2 Peer Selection

In every type of overlay network, a set of source peers is assigned to each participating peer. The main difference between tree-based and mesh-based overlays is the means to make this assignment. In tree-based overlays, peer selection is the outcome of the tree construction process. In mesh-based overlays, peers have a more active role in peer selection. Each peer can select which peers to downstream from and which peers to upstream to. Peer selection is important for mesh-based overlays to achieve consistency in topology as tree-based ones.

In CoolStreaming/DONet [26], the notions of *membership* and *partnership* are introduced. Each peer is considered as a member of the overlay and maintains a partial view of the currently connected peers. This is realized by means of a membership cache called *mCache* which contains a partial list of the unique

identifiers of the peers in the system. Peers establish partnership with some of their members in order to exchange video content.

When a new peer joins the system, it contacts a bootstrap peer. The bootstrap peer contains information about all the connected peers in its mCache. It selects a random set of the connected peers and hands it over to the new peer. The new peer populates its mCache with the provided information. Then it selects a random subset of the mCache to establish partnership with.

In order to maintain its mCache, its peer periodically generates *membership messages* which are distributed via the Scalable Gossip Management Protocol (SCAM). Each membership message contains four elements: (a) a sequence number, (b) the peer identifier, (c) the current number of partners, and (d) a time-to-live values which is the validity period of the message. When a peer receives a membership message with a new sequence number, it checks whether there is an entry in its mCache with the same peer identifier as the one in the message. Entries in mCache contain the same fields as the membership messages plus an extra field which contains the last update time. When an entry is created or updated, the last update time is updated accordingly. Changes in mCache may also occur upon establishment of partnership between two peers. In this case, partners exchange mCaches and they update their local entries.

Peer selection for video streaming can be modelled as a free-market resource economy problem. A serving peer can charge for the delivery of video content a price that may depend on: (a) the popularity of the content, (b) the transfer rate, and (c) the duration of transfer. Price is given by pricing functions of the form $c_i(b_i) \cdot t_i$, where b_i is the transfer rate from server i, t_i is the duration of transfer, and $c(.)$ is a cost-rate function. The cost-rate function is equivalent to the cost per unit time when the server streams at rate b.

Assume that a video object has a playback rate r and that the viewing time is T seconds. A peer can receive the content from a set $\{1, \ldots, I\}$ of server peers. When server peer i streams at rate b_i, the cost per unit time is $c_i(b_i)$. The receiving peer must select the portion of the video object that will download from each server and also the download rate. The division in video portions can be realized in the time domain or in the rate domain. In time domain, the video object is partitioned in the time axis and each server peer streams one partition. The peer downloads different partitions in parallel and the aggregate rate varies as download of different partitions terminates. In rate domain, each server peer streams a portion of each video frame and the peer downloads at a rate that equals to the playback rate.

In [1], the problem is formulated in rate domain as follows: The peer must choose a set of server rates $b_1, \ldots, b_I$ such that the total cost $c_1(b_1)T + \ldots + c_I(b_I)T$ is minimized subject to the constraint that the peer must receive at least at rate r during viewing time T. The system must withstand a failure of a single server peer, namely j.

$$\min_{\mathbf{b}} \sum_{i=1}^{I} c_i(b_i) \tag{8.1}$$

$$\text{s.t.} \sum_{i \neq j} b_i \geq r,\ j = 1, \ldots, I \tag{8.2}$$

$$0 \leq b_i \leq r,\ i = 1, \ldots, I \tag{8.3}$$

In case $c(.)$ is convex, the previous problem can be solved by means of the following sub-problem, for any given $y : 0 \leq y < r$:

$$\min_{\mathbf{b}} \sum_{i=1}^{I} c_i(b_i) \tag{8.4}$$

$$\text{s.t.} \sum_{i} b_i \geq r + y,\ j = 1, \ldots, I \tag{8.5}$$

$$0 \leq b_i \leq y,\ i = 1, \ldots, I \tag{8.6}$$

The sub-problem stated in (8.4)–(8.6) is solved by a marginal allocation algorithm [1].

Algorithm 5 Marginal allocation algorithm for problem (8.4)–(8.6)

$\mathcal{S} := \{1, \ldots, I\}$
$b_i = 0\ \forall i \in \mathcal{S}$
repeat
  $i^* = \arg \min_{i \in \mathcal{S}} \{c_i(b_i + \Delta) - c_i(b_i)\}$
  $b_{i^*} \leftarrow b_{i^*} + \Delta$
  if $b_i > y - \Delta$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} - i$
  end if
until $\sum_i b_i \geq r + y$

The original problem (8.1)–(8.3) is then solved with a greedy algorithm.

Algorithm 6 Algorithm for solving problem (8.1)–(8.3)

$y^* = r/(I - 1)$
repeat
  solve problem (8.4)–(8.6) with for $y = y^*$
  find solution $C(y)$
  $y^* \leftarrow y^* + \Delta$
until $C(y)$ does not decrease **or** $y = r$

The previous resource-economy-based approach does not guarantee efficient use of network resources. Instead, it guarantees cost-effective use of resources given the fact that each peer defines the cost-rate function of its resources arbitrarily.

Alternatively, peer selection can be modelled as a cooperative game among peers as presented in [25]. In the simple case with a single parent, a parent p and a set of children $c_1, c_2, \ldots, c_n$ are players and they can form *coalitions*. A coalition can be the union of the parent with any subset of children set. The problem is to find a stable coalition with the highest aggregate value.

Each peer x that participates in a coalition G gains a value $v(x)$. The value $V(G)$ of coalition G is the sum of the values allocated to all the peers:

$$V(G) = \sum_{\forall x \in G} v(x) \tag{8.7}$$

The coalition is stable if peers have no incentive to decide not to join the coalition. In this case, the coalition G has higher value than any other coalition $G' : G' \subseteq G$. A coalition G must satisfy the following conditions:

$$V(G) = 0 \; if \; p \notin G \tag{8.8}$$

$$V(G) \leq V(G') \; if \; G \subseteq G' \tag{8.9}$$

$$V(G_1 \cup c_i) - V(G_1) \neq V(G_2 \cup c_i) - V(G_2) \tag{8.10}$$

Equation (8.8) states that parent p must participate in any coalition. Equation (8.9) states that the value of a coalition cannot be lower than the value of any smaller coalition. Equation (8.10) states that the same peer has different contribution to the value of different coalitions. In order to participate in a coalition, a peer x puts a *coalitional effort* $e(x)$. The utility gained by the player is $u(x) = v(x) - e(x)$. The coalitional effort can be modelled as follows:

$$e(x) = \begin{cases} (|G| - 1)e & , \; x = p \\ e & , \; x \in G \setminus p \end{cases}$$

This is justified by the fact that the server has to put effort for each participating peer, whereas the effort of a peer is a constant parameter e. In [25], the following value function is proposed:

$$V(G) = \begin{cases} \log\left(1 + \sum_{\forall i \neq p} \frac{1}{b_i}\right), & p \in G \\ 0, & \text{otherwise} \end{cases}$$

Algorithm 7 Game-theoretic peer selection: parent side

loop
  upon receiving a request from potential child c_x
  calculate $v(c_x) = V(G_Y \cup c_x) - V(G_Y) - e$
  if $v(c_x) \geq e$ **then**
    reply with bandwidth allocation $b_{x,\,y} = \alpha \cdot v(c_x)$
  else
    reply with bandwidth allocation $b_{x,\,y} = 0$
  end if
end loop

Algorithm 8 Game-theoretic peer selection: child side

obtain m candidate parents from server
for each candidate parent y **do**
  send request to candidate parent y
  receive reply with bandwidth allocation $b_{x,\,y}$
end for
$B' = b_{x,y} \forall y; B =; b = 0;$
while b¡1 **do**
  select the largest allocation from $b_{x,\,y'}$ from B
  $B = B \cup y'$;
  $b = b + b_{x,y'}$;
end while
cancel allocation for parents not listed in B
confirm allocation for parents listed in B

In mesh-based overlays, peer selection must protect the connectivity of the overlay network. This means that each peer must have a minimum number of children peers.

Peer selection is more complex in wireless networks. Authors in [8] formulate the problem by means of network utility maximizations (NUM). In [3], the problem of peer selection for scalable video is addressed. Let $\mathscr{R}$ be the set of receiving peers in the network. Each receiving peer i has a different set $\mathscr{S}_i$ of $S(i)$ possible sending peers. Let $L(i)$ be the number of layers that each receiving peer i can decode and let $\mathscr{L}_i$ be the corresponding set of layers. The outcome of the optimization problem is the derivation of the decision matrices $x^{(i)} \forall i \in \mathscr{R}$. Each element $x_{sl}{}^{(i)}$ of the $S_i \times R_i$ decision matrix $x^{(i)}$ takes binary values and denotes whether receiving peer i receives layer l from sending peer s (value 1) or not (value 0).

We denote $b_l{}^i$ as the source bitrate of the lth layer of the video sequence requested by peer i. t_{mn} denotes the achievable throughput in the link (m, n) and t_{si} denotes the end-to-end throughput from peer s to peer i. Let h_{mn}^{si} be a binary

variable that describes whether link (m, n) belongs to the route between sending peer s and receiving peer i. We denote $\mathscr{H}$ the set of physically connected nodes, i.e. the set of one-hop neighbors.

The following interference model is used: A directed link between a sending node and a receiving node interferes with a set of other links if the receiving node is within the range of the sending nodes of the other links. Two interfering links should not be active at the same time. We adopt the notion of *maximal cliques* which is widely used in NUM formulations [14].

Given an undirected graph that depicts the physical network topology, an undirected *conflict graph* consists of a set of vertices that correspond to the directed physical links and a set of edges that indicate whether two physical links interfere with each other. Maximal cliques are maximal—in terms of number of vertices—subgraphs of the conflict graph in which each vertex is connected with every other vertex in the subgraph. Therefore, the complete set of maximal cliques $\mathscr{C}$ describes all possible interference scenarios. Let c_{mn}^{k} be a binary variable that equals 1 when link (m, n) is a vertex of maximal clique k. Due to interference, a link cannot be active constantly. Thus, we introduce the real variable y_{mn} to express the percentage of time link (m, n) is scheduled to be active.

Due to imperfect link scheduling and failures of the collision detection/avoidance mechanisms, links that belong to a maximal clique cannot be fully utilized. To capture this phenomenon, we introduce the link utilization factor a_{f}.

According to the assumptions described previously, we define the following optimization problem:

$$\underset{\mathbf{x},\mathbf{y}}{\text{maximize}} \sum_{i=1}^{R} U_i(x^{(i)}) \tag{8.11}$$

$$\text{s.t.} \sum_{s=1}^{S(i)} x_{sl}^{(i)} \leq 1,\ \forall i \in \mathscr{R}, l \in \mathscr{L}_i \tag{8.12}$$

$$\sum_{s=1}^{S(i)} x_{sl_1}^{(i)} \geq \sum_{s=1}^{S(i)} x_{sl_2}^{(i)},\ \forall i \in \mathscr{R}, l_1, l_2 \in \mathscr{L}_i : l_1 < l_2 \tag{8.13}$$

$$\sum_{i=1}^{R} \sum_{s=1}^{S(i)} \left[h_{mn}^{si} \sum_{l=1}^{L(i)} (x_{sl}^{(i)} b_l^i) \right] \leq t_{mn} y_{mn},\ \forall (m,n) \in \mathscr{H} \tag{8.14}$$

$$\sum_{m} \sum_{n} c_{mn}^{k} y_{mn} \leq a_{\mathrm{f}},\ \forall (m,n) \in \mathscr{H}, k \in \mathscr{C} \tag{8.15}$$

$$\sum_{l=1}^{L(i)} x_{sl}^{(i)} b_l^i \leq t_{si},\ \forall i \in \mathscr{R}, s \in \mathscr{S}_i \tag{8.16}$$

$$x_{sl}^{(i)} \in \{0, 1\},\ \forall i \in \mathscr{R}, s \in \mathscr{S}_i, l \in \mathscr{L}_i \tag{8.17}$$

$$0 \leq y_{mn} \leq 1,\ \forall (m,n) \in \mathscr{H} \tag{8.18}$$

The objective of the problem is to maximize the network utility. As shown in (8.11), the network utility is the sum of the individual utility functions U_i of the receiving peers. Furthermore, U_i is a function of the decision matrix $x^{(i)}$. In general, each peer wishes to receive as many quality layers as possible, given congestion and interference constraints derived from the network topology. In addition, lower quality layers are more important than higher quality layers. Since the elements $x_{sl}^{(i)}$ of the decision matrices are binary and since a layer is received by a single sending peer, the sum $\sum_{s=1}^{S(i)} x_{sl}^{(i)}$ represents whether peer i receives layer l, regardless from which is the sending peer. The sum $\sum_{l=1}^{L(i)} \sum_{s=1}^{S(i)} x_{sl}^{(i)}$ then represents the number of layers received by peer i.

To capture the different importance of different layers, we assign a real positive weight factor w_l to each layer l. Then we define the following utility function:

$$U_i(x^{(i)}) = \sum_{l=1}^{L(i)} \left[w_l \left(\sum_{s=1}^{S(i)} x_{sl}^{(i)} \right) \right] \tag{8.19}$$

$U_i(x^{(i)})$ is an increasing function of the number of admissible layers. Since lower layers are more important in the decoding process, it is rational to introduce the following inequality:

$$w_l > \sum_{j=l+1}^{L(i)} w_j, \ \forall i \in \mathscr{R}, l \in [1, L(i) - 1] \tag{8.20}$$

Equation (8.20) states that admitting a single layer l is more important than admitting all the layers that are higher than l. The choice of weight factors that satisfy (8.20) is also a means for providing fairness among flows of different video sequences. Inequality (8.12) expresses that a layer is exclusively sent by a single source. Inequality (8.13) expresses the inter-layer dependency constraint. It indicates that a layer can be admitted if and only if its directly lower layer can also be admitted. Inequality (8.14) is the congestion constraint. The amount of video traffic that traverses a link must not exceed the link throughput multiplied by the link schedule. If we wish the proposed scheme to be friendly to existing traffic in the network, (8.14) can be replaced by:

$$\sum_{i=1}^{R} \sum_{s=1}^{S(i)} \left[h_{mn}^{si} \sum_{l=1}^{L(i)} (x_{sl}^{(i)} b_l^i) \right] \leq y_{mn} \left[t_{mn} - \sum_{f=1}^{F} (h_{mn}^f b^f) \right], \ \forall (m,n) \in \mathscr{H} \tag{8.21}$$

In (8.21), F is the number of non-video flows in the network, $h_{mn}{}^f$ indicates whether link (m, n) is used by flow f, and b^f is the bitrate of flow f. Inequality (8.15) represents the interference constraint. Inequality (8.16) is used to ensure that a video flow does not exceed the end-to-end throughput between the sending peer and the receiving peer. This constraint accounts for the problem of throughput

Table 8.1 Notations for the peer selection problem in a wireless network

Notation	Description
$\mathscr{R}$:	Set of receiving peers
$\mathscr{S}_i$:	Set of possible sending peers for peer i
$S(i)$:	Number of possible sending peers for peer i
$\mathscr{L}_i$:	Set of layers for peer i
$L(i)$:	Number of layers for peer i
$x^{(i)}$:	Decision matrix for peer i
$x_{sl}^{(i)}$:	1 if peer i receives layer l from peer s
b_l^i:	Source bitrate of layer l
t_{mn}:	Achievable throughput in the link (m, n)
t_{si}:	End-to-end throughput from peer s to peer i
h_{mn}^{si}:	1 if link (m, n) belongs to the route from peer s to peer i
$\mathscr{H}$:	Set of physically connected nodes
$\mathscr{C}$:	Set of maximal cliques
c_{mn}^k:	1 if link (m, n) is a vertex of maximal clique k
y_{mn}:	Percentage of time link (m, n) is active
a_f:	Link utilization factor
$U_i()$:	Utility function of peer i
w_l:	Weigh factor of layer l

degradation which is observed in multihop wireless networks. Finally, constraint (8.17) forces the decision variables $x_{s\ l}^{(i)}$ to take only binary values and constraint (8.18) limits value field of link schedule $y_{m\ n}$ between 0 and 100 %. Table 8.1 summarizes the notation used in the formulation of the problem.

It is noted that although $y_{m\ n}$ are not included in the network utility, feasible link schedules must also be computed in order to produce an optimal peer selection and layer allocation (both expressed by the binary decision variables $x_{sl}^{(i)}$). Therefore, (8.11)–(8.18) constitute an MILP. We can relax the constraint (8.17) and allow $x_{sl}^{(i)}$ to take real values.

$$x_{sl}^{(i)} \in [0, 1],\ \forall i \in \mathscr{R}, s \in \mathcal{S}_i, l \in \mathscr{L}_i \tag{8.22}$$

However, in this case (8.12) does not guarantee single source transmission per layer. Instead, the transmission of any layer is possible to be assigned to several sending peers, each contributing a different percentage. Equation (8.12) only guarantees that no redundant transmissions will occur. In order to force the system allocate complete layers to sending peers, we adopt a new utility function. Instead of using the weighted sum of the number of layers, we use the weighted sum of the square of the values of the decision variables.

$$U_i'(x^{(i)}) = \sum_{l=1}^{L(i)} \left[w_l \sum_{s=1}^{S(i)} \left(x_{sl}^{(i)} \right)^2 \right] \tag{8.23}$$

The problem is then transformed to an NLP. The coupling constraints (8.14) and (8.15) can be relaxed via Lagrangian multipliers. Then, the resulting sub-problems are charged with updating the primal variables $x_{sl}^{(i)}$ and $y_{m\ n}$ while the master problem updates the values of the Lagrangian multipliers. The solution is sub-optimal with respect to the initial MILP problem.

The problem of scheduling transmissions from multiple senders in wireless peer-to-peer networks is addressed in [18] and [19]. The proposed mechanism is based on the *multi-armed bandit* approach and tries to: (a) maximize the receiving data rate and (b) minimize power consumption. According to this approach, selecting a sender at each time slot gives a certain reward value. The aim is to define a sender schedule in order to maximize the total reward. The problem is based on a partially observed Markov decision process (POMDP).

Among multiple servers, the one with the largest *Gittins index* acts as the *active sender*. The Gittins index $\gamma^k(l, x^k(l))$ is a function of the sender l and its *information state* $x^k(l)$ at time slot k. A sender l can be in U_l different states. We denote by $s^k(l)$ the state of sender l at time slot k. The information state $x^k(l)$ is defined as the following matrix:

$$x^k(l) = (x_i^k(l)), i = 1, 2, \ldots, U_l \tag{8.24}$$

Each element $x_i^k(l)$ is defined as follows:

$$x_i^k(l) = \Pr[s^k(l) = i | Y^k, a^{k-1} = l] \tag{8.25}$$

where Y^k is the *observation history* and a^{k-1} is the active sender at time slot $k - 1$. Algorithm 9 describes the sender scheduling process.

Algorithm 9 Distributed sender scheduling

Off-line computation:
for each sender $l = 1, 2, \ldots, L$ **do**
    Compute a finite set of vectors $\Lambda^N(l)$ and store them in the index tables.
end for
Compute $\gamma^K(l, x^0(l))$.
Real-time scheduling:
Find the address list of the L potential senders.
At $k = 0$, send a request for the first chunk to all potential servers and include the address list in the request message.
Each sender decides $\gamma^K(l, x^0(l))$ and multicasts the index to other senders.
repeat

(continued)

Algorithm 9 (continued)

Each sender stores the L-dimensional vector (a^k, γ) where a^k is the active sender and γ is the vector of the Gittins indices of the L senders sorted in descending order.
Sender 1 transmits the requested chunk to the receiver.
At the next time slot, sender 1 obtains the observation $y^{k+1}(1)$, updates the state estimation and decides it $\gamma^K(l, x^{k+1}(l))$.
Keep the Gittins indices unchanged for the remaining senders.
if $\gamma^K(l, x^{k+1}(l)) \geq \gamma^K(l, x^k(l))$ **then**
Sender 1 will continue to be active.
else
Sender 1 multicasts $\gamma^K(l, x^{k+1}(l))$ to other potential senders and becomes passive.
end if
until The last chunk is successfully transmitted.

8.3.3 Chunk Scheduling

In mesh-based overlays following the pull-based approach, each peer receives requests from multiple peers in order to provide them with specific chunks. Each peer must determine which peers will be served first and also the allocation of the available upload bandwidth to the requesting peers. Reversely, peers that request chunks must determine which chunks to request, from which peer to request each chunk, and at what order to issue the requests. The main factors that affect these decisions are: the rareness of the chunks, the playback deadlines, and the relative importance of chunks.

BitTorrent adopts a *rarest-first* policy for chunk selection, i.e. chunks with less potential suppliers are requested first. The same policy is proposed in [17] with the following modification: The size w of the sliding window must be adjusted so that a peer can use the playback delay d to download all necessary chunks to play the first d time units of the stream. Given the video consumption rate b and the chunk size c, the size of the sliding window must satisfy the following relation:

$$w = \frac{db}{c} \tag{8.26}$$

BiToS [21] is another system inspired by BitTorrent. BiToS enhances BitTorrent by adding a *view-as-you-download* service. Three additional components enable smooth streaming performance, namely *Received Pieces Set*, *High Priority Set*, and *Remaining Pieces Set*. The Received Pieces Set contains all the downloaded chunks. Chunks in this set can be of the following types: *Downloaded*, *Not-downloaded*, and *Missed*. The High Priority Set contains all the chunks that have not been Downloaded yet, are not Missed and their play-out deadline is approaching. Chunks in this set can be of the following types: *Not-Requested* and

Currently-Downloading. Finally, the Remaining Pieces Set contains all the chunks that have not been Downloaded, are not Missed and are not in the Priority Set. A chunk in the Remaining Pieces Set can be *Not-Requested* or *Currently-Downloading*. The chunk scheduling algorithm in BiToS is described next.

Each time a peer selects a chunk for downloading, there is a probability p that the selected chunk is picked from the High Priority Set and a probability $1 - p$ that the chunk is picked from the Remaining Pieces Set. The value of p represents a trade-off between playback continuity and acquisition of chunks that will be needed in the future. Chunks from the High Priority Set are chosen according to the Rarest-First mechanism. However, if some chunks have the same rareness, the chunk which is closer to meet its play-out deadline is selected. At any time, there is maximum of k Currently-Downloading chunks.

When a chunk download is complete, the chunk is moved from its set (High Priority Set or Remaining Pieces Set) to the Received Pieces Set. If the downloaded chunk was in the High Priority Set, a chunk from the Remaining Pieces Set is moved to the High Priority Set. This chunk is the one with the closest play-out deadline. A peer can provide other peers with Downloaded chunks in its Received Pieces Set. Algorithm 10 presents the chunk scheduling in BiToS.

Algorithm 10 Chunk scheduling algorithm in BiToS

```
loop
    Choose the set to download the next chunk from:
    Prob. p: High Priority Set or Prob. (1 − p): Remaining Pieces Set
    Find rarest chunks within the selected set
    if There are many chunks with the same rareness then
      Select the chunk with the closest deadline within the rarest chunks
    else
      Select the rarest chunk
    end if
    Download the selected chunk from the selected set
    Include chunk in the Received Pieces Set
    if Download was successful and the chunk was received on time then
      Mark chunk as 'Downloaded'
    else if Download was unsuccessful then
      Mark chunk as 'Not-Downloaded'
    else if Download was successful but the chunk was not received on time then
      Mark chunk as 'Missed'
    end if
    if The selected set was in the High Priority Set then
      Find the chunk with the closest deadline in the Remaining Pieces Set and
      move it to the High Priority Set
    end if
    Proceed to next chunk
end loop
```

Table 8.2 Parameters for the chunk scheduling algorithm in Coolstreaming/DONet

Notation	Description
band[k]:	Bandwidth from partner k
bm[k]:	Buffer map of partner k
bm[j, i]:	1 if partner j has chunk i in its buffer map
t[j, i]:	Available time for transmitting chunks until i
deadline[i]:	Deadline of chunk i
chunk_size:	Segment size
num_partners:	Number of partners
set_partners:	Set of partners
expected_set:	Set of chunks to be fetched
chunk_sets[n]:	Sets of chunks with n suppliers

In CoolStreaming/DONet [26], the chunk scheduling algorithm tries to meet two constraints: the playback deadline for each chunk and the heterogeneous bandwidth from different partners. Since finding optimal schedules that meet these constraints is NP-hard, a heuristic method is used. According to this method, chunks are scheduled to be fetched according to the number of potential suppliers. A potential supplier of chunk c for peer p is any partner of p that possesses chunk c. Chunks with less potential suppliers are served first. For each chunk, the peer with the highest bandwidth and enough available time is selected as supplier.

The algorithm runs on a periodical basis. At the beginning of each period, each peer tries to schedule the delivery of a set of expected chunks called *expected set* (denoted as expected_set). Prior to the execution of the algorithm, partners exchange their buffer maps. Therefore, each peer has the buffer map bm[k] of every peer k in its set of partners (set_partners). A binary element bm[j, i] in the buffer map bm[j] declares whether partner j has chunk i in its buffer. At first, the available times t[j, i] for transmitting chunks until chunk i and the number n of potential suppliers are computed for each chunk i. In addition, chunks are categorized according to the number of potential suppliers. Table chunk_sets[n] includes chunks with n potential suppliers. If the chunk i has only one potential supplier ($n = 1$), then this supplier k is selected and its available time is reduced by the time to transmit the chunk. The transmission time is the rate of the chunk size chunk$_s$ize to the bandwidth band[k] from partner k. After that, chunks with more than one potential suppliers are served. Algorithm 11 describes the previous process in detail and Table 8.2 presents the involved parameters.

Algorithm 11 Chunk scheduling algorithm in Coolstreaming/DONet

loop
  for segment $i \in$ *expected_set* **do**
    $n \leftarrow 0$
    for j **to** *num_partners* **do**

(continued)

Algorithm 11 (continued)

```
        t[i, j] ← deadline[i] − current_time
        n ← n + bm[j, i]
      end for
      if n = 1 then
        k ← arg_r{b m[r, i] = 1}
        supplier[i] ← k
        for j ∈ expected_set, j > k do
  t[k,j] ← t[k,j] − seg_size/band[k]
        end for
      else
        chunk_sets[n] ← dup_set[n] ∪{ i}
        supplier[n] ← null
      end if
    end for
    for n = 2 to num_partner s do
      for each i ∈ chunk_sets[n] do
                k ← arg_r {band(r) > band(r′)|t[r, i] > chunk_size/band[r], t[r′, i]
> chunk_size/band[r′], r, r′ ∈ set_partners}
        if k≠null then
  supplier[i] ← k
  for j ∈ expected_set, j > k do
  t[k,j] ← t[k,j] − chunk_size/band[k]
  end for
        end if
      end for
    end for
  end loop
```

According to the approach described in [10], each source peer randomly selects which receiving peer to serve. The selection is performed among neighbors which have pending requests. Within the mesh-based approach, two neighboring peers can have a two-way source–receiver relation. The source peer favors neighbors from which it downloads chunks at a higher rate. The probability $p_{n,\,k}$ that source peer n selects receiving peer k is given by:

$$p_{n,k} = \frac{I_{n,k}(d_{n,k} + \epsilon)}{\sum_{i \in \mathscr{K}_n} I_{n,i}(d_{n,i} + \epsilon)} \tag{8.27}$$

where $\mathscr{K}_n$ is peer n's set of neighbors, $d_{n,\,i}$ is the uploading rate from peer i towards peer n, and ϵ is a small positive constant. $I_{n,\,i}$ and $I_{n,\,k}$ are binary variables that express whether there are pending requests to peer n from peers i and k, respectively. The role of ϵ is to ensure that a peer will be considered in the selection process even if it has not contributed its uploading bandwidth yet.

A receiving peer with low uploading rate towards the source peer will not necessarily be rewarded with low uploading rate by the source peer. If the network is under-loaded, there will be fewer pending requests and therefore chunks requested by peers with low $d_{n,i}$ will have increased chances to be served.

Apart from the chunk selection and bandwidth allocation exercised by the source peer, receiving peers must also select source peers and schedule requests for chunks properly. A common approach relies on random scheduling. Each receiver periodically requests missing chunks. Chunks are requested randomly, without any form of prioritization. If a missing chunk is available from multiple providers, one of them is chosen at random.

Authors in [10] propose a scheduling algorithm at the receiver side which is suitable for handling layered video streaming. Requests for important chunks are expected to be served on time whereas requests for less important chunks are expected to be served if the corresponding source peers have enough upload bandwidth. The former are called *regular requests* and the latter are called *probing requests*. There is a layer threshold l_n below which requests are regular.

The proposed algorithm computes the chunk schedule according to the following input variables: the set $\mathscr{K}_n$ of neighbors of peer n, the set $\mathscr{C}_n$ of chunks to be scheduled, the number of layers for regular requests l_n, and the total number of layers L. Let $L_{i\ d}[i]$ denote the layer index of chunk i.

Algorithm 12 Chunk scheduling algorithm at the receiver

loop
for $i \in \mathscr{C}_n \wedge L_{id}[i] \leq l_n$ **do**
for $k \in \mathscr{K}_n$ **do**
if chunk i is owned by neighbour k **then**
insert k to supplier set
end if
end forrandomly select a neighbour k^* from the supplier set to request chunk i from
end for
for $l \leftarrow l_n + 1$ to L **do**
for $i \in \mathscr{C}_n \wedge L_{id} = l$ **do**
for $K \in \mathscr{K}_n$ **do**
if chunk i is owned by neighbour k **then**
insert k to supplier set
end if
end for
randomly select a neighbour k^* from the supplier set to request chunk i from
end for
end for
end loop

An important goal of chunk scheduling is the efficient use of the system resources in favor of streaming performance. The scalability of a peer-to-peer system depends on the exploitation of the upload bandwidth of peers. As suggested in [7], proper chunk scheduling can achieve almost full utilization of the upload bandwidth.

In this system, there are source servers and peers. Each peer has two queues: a *playback buffer* and a *forwarding queue*. The playback buffer stores received chunks from different peers in playback order. The forwarding queue contains chunks to be forwarded to other peers. Incoming chunks are divided into two classes: F chunks and $N\,F$ chunks. F chunks are consumed by the peer and they are forwarded to other peers, whereas $N\,F$ chunks are consumed without being forwarded. Only chunks received by the source server are marked as F. In order to achieve full utilization of the peers' upload bandwidth, the forwarding queue must be always busy. Whenever the forwarding queue is empty, the peer sends a *pull* message to the server in order to request more chunks.

The server has also three queues: a *content queue*, a *signal queue*, and a *forwarding queue*. The content queue stores the chunks to be forwarded to peers. It has two dispatchers: a *content dispatcher* and a *forward dispatcher*. The signal queue stores signals received from peers. If there is a "pull" signal in the signal queue, a chunk is taken from the content queue, it is marked as F and it is dispatched by the content dispatcher. The content dispatcher forwards it to the peer that issued the "pull" signal and the "pull" signal is removed from the signal queue. If the signal queue is empty, a chunk is taken from the content queue, it is marked as $N\,F$ and it is dispatched by the forward dispatcher. The forward dispatcher puts that chunk to the forwarding queue to be forwarded to all the peers.

Theoretically, the chunk scheduling practice described previously achieves maximum streaming rate provided that the propagation delay is negligible and chunks can be arbitrarily small. In a real system, peers must adjust the timing of "pull" signals and the server must increase the number of chunks dispatched upon reception of a "pull" signal.

The peer sends a "pull" signal when the number of chunks in its content queue is less than or equal to a threshold T_i. The server sends K chunks when it dispatches a "pull" signal from a peer. In order to fully utilize a peer's upload bandwidth, its forwarding queue must always be busy. In this case, the following rule holds [7]:

$$T_i \geq \frac{(2t_{si} + K\delta)/u_s + t_q)u_i}{(n-1)\delta} \tag{8.28}$$

where $t_{s\ i}$ denotes the propagation delay between server s and peer i, u_s denotes the upload capacity of the server, u_i denotes the upload capacity of the peer, δ denotes the chunk size, and t_q denotes the queueing delay in the signal buffer at the server.

8.4 Comparison Between Tree-Based and Mesh-Based Overlays

Tree-based and mesh-based overlays for peer-to-peer streaming are characterized by fundamental similarities and differences as well. The following similarities hold [11]:

1. *Similar structure*: Regardless of the means of overlay formation and maintenance, both tree-based and mesh-based overlays are characterized by a tree-like structure.
2. *Similar content delivery*: In both approaches, peers receive different pieces of video content from different parents and forward these pieces to children peers.
3. *Need for loosely synchronized play-out buffer*: In tree-based overlays, buffering is essential in order to accommodate delay diversity among flows traversing different trees. In mesh-based systems, buffering is required to accommodate chunks that are received at random order and also to schedule requests for missing chunks.

Tree-based and mesh-based systems differ in the following items [2, 11]:

1. *Formation of delivery trees*: In tree-based systems, the delivery trees are static whereas in mesh-based systems the delivery trees are formed dynamically as an indirect outcome of the peer selection process.
2. *Partitioning of the video content*: Due to the dynamic nature of mesh-based systems, partitioning of video content into chunks is necessary. On the contrary, in tree-based systems, the partitioning of the video flow into sub-flows and the allocation of the video sub-flows to the sub-trees are performed at packet level.
3. *Impact of delay*: In tree-based systems, all packets traversing a logical link are subject to the transmission delay of this logical link which is the sum of the underlying link delays. However, in mesh-based systems, the transmission delay is not the only source of delay and the available bandwidth plays a significant role in delay performance. These phenomena are elaborated next.

In tree-based overlays, optimization of delay performance is a minimum path cost problem. Specifically in case of homogeneous links, the problem is reduced to minimizing the depth of the trees. Given that there is enough upload bandwidth for a peer to support its outbound connections, any further increase in the upload bandwidth does not improve delay performance.

In mesh-based systems, transmission delay plays a more important role because chunks are disseminated in a store-and-forward manner. A peer cannot start forwarding a chunk to other peers until it has finished receiving this chunk from its parent peer. Therefore, peers themselves increase delay when acting as relays. In addition, minimization of delay cannot be formed as a minimum path cost problem for a series of reasons analyzed in [2]. At first, the delay observed during the transmission of a single chunk depends on the fraction of the upload bandwidth that is dedicated to this transmission. Secondly, after a peer has received a chunk, it forwards this chunk to a set of peers in a sequential manner. Consequently, the

delay until all these children peers obtain the chunk is considerable. Thirdly, a peer must accomplish the transfer of a chunk to a set of children peers before a new chunk has been generated. Therefore, in mesh-based systems, the number of children peers is constrained.

In [22, 23], a hybrid system that combines tree-based and mesh-based approaches is proposed. This system, called *mTreebone* constructs a tree-based backbone of stable peers. Stable peers also participate in auxiliary mesh-based sub-overlays.

8.5 Conclusions

Peer-to-peer streaming represents a promising technology that enables the delivery of video content to thousands of users in a scalable and efficient manner. Two main types of peer-to-peer streaming networks exist, namely tree-based and mesh-based.

Tree-based networks rely on application-layer multicast trees. In each tree, each peer receives content from a parent peer and forwards it to a set of children peers. The main research issue in tree-based systems is the overlay formation and maintenance. More specifically, the structure of the trees must withstand peer dynamics and exploit peers available resources effectively. Tree-based systems perform well in terms of delay because of the predefined topology. On the other hand, they are prone to churn because the overlay has to be repaired after peer arrival and departures.

Mesh-based networks do not rely on a predefined topology. Instead, peers form partnerships and periodically exchange information about the availability of video chunks. In push-based systems, each peer provides other peers with new chunks. In pull-based systems, each peer requests certain chunks from other peers. Pull-based systems represent the dominant approach both in academic research and in implemented systems. Two main issues are present in mesh-based systems: peer selection and chunk scheduling. Peer selection is defined as the process of selecting peers to establish partnerships with. Chunk scheduling is defined as the process of scheduling pull requests to other peers in order to obtain missing chunks. Mesh-based systems can tolerate frequent peer arrival and departures. However, they perform poorly in terms of delay.

Recent research efforts have shed light on several aspects of peer-to-peer streaming. Practical algorithms as well as analytical tools have been proposed. Yet, there are still many open issues that need to be addressed. Given the increasing demand for peer-to-peer video content, peer-to-peer streaming represents an active research area with direct benefits for the end-users and content providers.

References

1. Adler M, Kumar R, Ross K, Rubenstein D, Suel T, Yao DD (2005) Optimal peer selection for p2p downloading and streaming. In: Proceedings of IEEE INFOCOM 2005. 24th Annual joint conference of the IEEE computer and communications societies, Miami, FL, U.S.A. vol 3, pp 1538–1549
2. Bianchi G, Melazzi NB (2009) Fundamental delay bounds in peer-to-peer chunk-based real-time streaming systems. In: Proceedings of 21th international teletraffic congress (ITC), Paris, France, pp 1–8
3. Birkos K, Tselios C, Dagiuklas T, Kotsopoulos S (2013) Peer selection and scheduling of H. 264 SVC video over wireless networks. In Wireless communications and networking conference (WCNC), 2013 IEEE, IEEE, Shanghai, China, pp 1633–1638
4. Castro M, Druschel P, Kermarrec A-M, Rowstron AIT (2002) Scribe: a large-scale and decentralized application-level multicast infrastructure. IEEE J Sel Areas Commun 20 (8):1489–1499
5. Castro M, Druschel P, Kermarrec A-M, Nandi A, Rowstron A, Singh A (2003) SplitStream: high-bandwidth content distribution in cooperative environments. In: Proceedings of ACM SOSP 2003, Sagamore, Bolton Landing, NY, U.S.A., pp 292–303
6. Chiu Y-M et al (2008) Minimizing file download time in stochastic peer-to-peer networks. IEEE/ACM Trans Netw 16(2):253–266
7. Guo Y, Liang C, Liu Y (2008) AQCS: adaptive queue-based chunk scheduling for p2p live streaming. In: NETWORKING 2008 Ad Hoc and sensor networks, wireless networks, next generation Internet. Springer, Berlin, pp 433–444
8. Gurses E, Kim AN (2008) Maximum utility peer selection for p2p streaming in wireless ad hoc networks. In: IEEE global telecommunications conference, 2008 (IEEE GLOBECOM 2008), New Orleans, LA, U.S.A., pp 1–5
9. Liu Y, Guo Y, Liang C (2008) A survey on peer-to-peer video streaming systems. Peer-to-peer Netw Appl 1(1):18–28
10. Liu Z, Shen Y, Ross KW, Panwar SS, Wang Y (2009) Layerp2p: using layered video chunks in p2p live streaming. IEEE Trans Multimed 11(7):1340–1352
11. Magharei N, Rejaie R (2007) Mesh or multiple-tree: a comparative study of live P2P streaming approaches. In: IEEE INFOCOM 2007. 26th IEEE international conference on computer communications, Anchorage, Alaska, U.S.A., pp 1424–1432
12. Padmanabhan VN (2003) Resilient peer-to-peer streaming. In: Proceedings of 11th IEEE international conference on network protocols, Austin, U.S.A., pp 16–27
13. Padmanabhan VN, Wang HJ, Chou PA, Sripanidkulchai K (2002) Distributing streaming media content using cooperative networking. In: Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video. Association for Computing Machinery, Miami, FL, U.S.A., pp 177–186
14. Palomar DP, Chiang M (2006) A tutorial on decomposition methods for network utility maximization. IEEE J Sel Areas Commun 24(8):1439–1451
15. Ramzan N, Park H, Izquierdo E (2012) Video streaming over P2P networks: challenges and opportunities. Signal Process Image Commun 27(5):401–411
16. Rowstron A, Druschel P (2001) Pastry: scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In: Middleware 2001. Springer, Berlin, pp 329–350
17. Shah P, Paris J-F (2007) Peer-to-peer multimedia streaming using BitTorrent. In: IEEE international performance, computing, and communications conference, 2007 (IPCCC 2007), New Orleans, U.S.A., pp 340–347
18. Si P, Yu FR, Ji H, Leung VCM (2009) Distributed sender scheduling for multimedia transmission in wireless mobile peer-to-peer networks. IEEE Wirel Commun 8(9):4594–4603
19. Si P, Yu F, Ji H, Leung VCM (2010) Distributed multisource transmission in wireless mobile peer-to-peer networks: a restless-bandit approach. IEEE Trans Vehicular Technol 59(1):420–430

20. Venkataraman V, Francis P (2006) Chunkyspread: multi-tree unstructured peer-to-peer multicast. In: Proceedings of the 2006 14th IEEE international conference on network protocols, (ICNP '06), Santa Barbara, CA, U.S.A., pp 2–11
21. Vlavianos A, Iliofotou M, Faloutsos M (2006) Bitos: enhancing BitTorrent for supporting streaming applications. In: Proceedings of INFOCOM 2006. 25th IEEE international conference on computer communications, Barcelona, Spain, pp 1–6
22. Wang F, Xiong Y, Liu J (2007) mTreebone: a hybrid tree/mesh overlay for application-layer live video multicast. In: Proceedings of the 27th international conference on distributed computing systems, Toronto, ON, Canada
23. Wang F, Xiong Y, Liu J (2010) mTreebone: a collaborative tree-mesh overlay network for multicast video streaming. IEEE Parallel Distrib Syst 21(3):379–392
24. Yang X, De Veciana G (2004) Service capacity of peer to peer networks. In: INFOCOM 2004. Twenty-third annual joint conference of the IEEE computer and communications societies, Hong Kong, pp 2242–2252
25. Yeung MKH, Kwok YK (2009) On game theoretic peer selection for resilient peer-to-peer media streaming. IEEE Trans Syst 20(10):1512–1525
26. Zhang X, Liu J, Li B, Yum Y-SP (2005) Coolstreaming/donet: a data-driven overlay network for peer-to-peer live media streaming. In: Proceedings of IEEE INFOCOM 2005. 24th Annual joint conference of the IEEE computer and communications societies, Miami, FL, U.S.A., vol 3, pp 2102–2111

Chapter 9
IP-Based Mobility Scheme Supporting 3D Video Streaming Services

Asimakis Lykourgiotis, Riccardo Bassoli, Hugo Marques, and Jonathan Rodriguez

Abstract Presently, a large proportion of the Internet traffic is multimedia streaming. Moreover, there has been a vast proliferation of multimedia-capable mobile devices equipped with multiple radio interfaces. In particular, it is foreseen that by 2016 video streaming will account close to 70 % of consumer mobile traffic. A major challenge for the future mobile Internet is the delivery of 3D multi-view video as it involves a large amount of data and is vulnerable to losses and end-to-end delay. Thus, it is of great importance to investigate the impact of IP mobility management to 3D video streaming. In this chapter, the mobility protocols proposed by Internet engineering task force to support IP Mobility will be presented and compared analytically. Moreover, the impact of the most well-known protocols in terms of PSNR on 3D Video will be assessed and evaluated through simulation. Additionally, as existing networks are becoming more and more heterogeneous the recently released IEEE 802.21 for media independent handover (MIH) will be presented. Finally this initial approach could be improved by adding network coding (NC) in order to correct errors and erasures. Thus a novel NC-MIH protocol based on subspace coding that guarantees reliability of transmissions at media independent handover function layer, independently from transport protocols or acknowledgement mechanisms in use will be introduced.

A. Lykourgiotis (✉)
Department of Electrical and Computer Engineering, University of Patras, Rio, Hellas
e-mail: asly@ece.upatras.gr

R. Bassoli • J. Rodriguez
Instituto de Telecomunicações, Campus Universitário Santiago, Aveiro, Portugal
e-mail: bassoli@av.it.pt; jonathan@av.it.pt

H. Marques
Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal
e-mail: hugo.marques@av.it.pt

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_9, © Springer Science+Business Media New York 2014

9.1 Introduction

Nowadays, a substantial portion of the Internet traffic is multimedia streaming. The vast proliferation of multimedia-capable mobile devices such as laptops and smart phones equipped with multiple radio interfaces (WLAN, 4G) intensifies the need for supporting IP mobility. However, the Internet was initially designed to facilitate fixed hosts for data exchange. The integration of complementary wireless technologies in order to achieve "anywhere, anytime, and on any device" ubiquitous services is the core idea of all-IP networks and the Future Internet. Through this concept, mobile nodes can roam or change network access technology continuing seamlessly their communication. Video is becoming a major challenge for the Future Internet as it will account close to 70 % of consumer mobile traffic by 2016 [4]. Among the wide variety of multimedia applications is the delivery of 3D multi-view video. It becomes apparent that 3D video streaming, especially due to the large amount of data involved and its time-sensitive nature, is a highly challenging problem for IP mobility management.

Mobility management is a major factor towards a seamless provision of multimedia applications in such a heterogeneous environment. Mobility management mainly comprises two tasks: location management and handover management. While location management is responsible for keeping track of the location of the Mobile Node (MN) between successive calls, handover management is in control of service continuity when the mobile node changes its point of attachment to the network. In order to support 3D video streaming, handover management must be extremely efficient with minimum handover delay. In this chapter, we survey different approaches for supporting IP Mobility that have been proposed in the leading standards development organization of Internet engineering task force (IETF). As the burdens imposed by live multimedia streaming were becoming more demanding innovative approaches were introduced to augment the previous protocols with new salient features. By investigating the evolution in mobility management protocols, the ability to deliver 3D video streaming is also under investigation.

Additionally, the IEEE 802.21 media independent handover (MIH) protocol will be presented. The 802.21 standard provides link layer intelligence and other related network information to upper layers to optimize handovers. In actual mobile networks, different technologies are coexisting and handsets with multiple wireless interfaces are moving under heterogeneous coverage. In this environment, there are two types of handovers: Horizontal and Vertical. A Horizontal Handover is a handover between two network access points (AP) that use the same network technology whereas a Vertical Handover (VHO) is a handover between APs of different network technologies. MIH, as it will be discussed, can assist vertical handovers between heterogeneous link layer technologies, as well as horizontal across IP subnet boundaries. In order to exchange messages among remote entities, the standard describes an MIH protocol. In case the link experiences errors and

erasures, the reliability of the protocol is guaranteed in two different ways either via a classical transport layer protocol such as transmission control protocol (TCP) or via the acknowledgement service of MIH protocol itself. According to these assumptions, it appears fundamental to find more efficient methods to augment MIH protocol strength against packet errors and erasure.

Network coding can represent an efficient solution to make MIH more reliable. In fact, network coding can also be useful for error correction. In particular, [5, 30] defined the capabilities and theoretic characteristics of network error correcting (NEC) codes. Then, subspace coding [11, 24] emerged as a new way of designing NEC codes, without the need of conveying the coefficients of the linear combinations into the header of the packets. This noncoherent approach communicates information via subspaces and exploits the fact that network linear operations do not affect them. Despite the amount of achievements in network coding theory, the research on network coding for practical application is still at the beginning: it is not yet clear how network coding can be developed into practice and the effects it can produce on existing systems

The remainder of this chapter is as follows. Initially, in Sect. 9.2 an overview of each proposed protocol is given considering the functions and entities that the protocol introduces as well as its handover signaling flow. Then, in Sect. 9.3 a novel NC-MIH protocol based on subspace coding will be described. Section 9.4 formulates a qualitative analysis to compare the previously described approaches and discusses which approaches are more suitable for efficient mobile video delivery. Additionally, in Sect. 9.5 simulation results will be presented to evaluate the impact of mobility protocols on 3D video streaming. Finally, Sect. 9.6 concludes the chapter.

9.2 IP Mobility Protocols Overview

In the following section the most well-known IP mobility management protocols will be presented. The protocols are categorized in global, local, route optimization enabled and fast handovers.

9.2.1 Global Mobility Protocols

Internet connections are set up and established utilizing the Sockets Application Programming Interface. Sockets uniquely identify the communications endpoints by an IP Address and a TCP/UDP port. Since the IP Address comprises a network prefix and a host identifier, the mobile node's change of subnet results to a prefix change which will destroy the socket and break any existing connections. A crucial concept in mobility management is the use of a unique identifier and a locator for the mobile node. An identifier is a stable value that uniquely identifies the mobile node,

regardless of its location while a locator is an IP address that indicates the mobile node's current attachment point to the Internet. As explained above applications use the IP address as an identifier of the host, while the Internet uses it as a locator. This duality of IP address is the reason why the Internet cannot inherently support mobility. Another important concept is the place where the mapping between mobile node's identifier and locator is held and it is named rendezvous point. In global mobility management the system is cognizant of the mobile node's current geographical location or subnet, i.e. its locator, on a global scale.

9.2.1.1 Mobile IPv4 (MIPv4)

MIPv4 [21] was introduced to alter the problem of socket close in TCP/IP and to allow IP to natively support mobility. As a result, MIPv4 allows the mobile node to utilize two addresses, one as an identifier called home address and the other as a locator named care-of-address. In order to enable the usage of these two addresses, new functions must be added to both the wireless routers and the mobile node. In MIPv4, the router that advertises a network having a prefix matching that of mobile node's home address is the rendezvous point called Home Agent (HA) and allows roaming of its users to other subnets. Additionally, a Foreign Agent (FA) is a router of any other network that is capable of accepting visitors. MIPv4 is a well-known protocol, sharing similarities with the next version and thus will not be further described herein.

9.2.1.2 Mobile IPv6 (MIPv6)

MIPv6 [23] provides a vast number of IP addresses as well as mechanisms that allow the mobile node to acquire its care-of address by itself, eliminating the need of foreign agents deployment. The two procedures that allow the mobile node to handle each own mobility management are the Router Discovery (RD) and the duplicate address detection (DAD). The former involves exchange of messages called Router Solicitation (RtSol) and Router Advertisements (RtAdv). The router advertisement message allows the mobile node to learn the link-local address of the access router (AR), the network prefixes, and the type of address configuration to use. As defined in [20], router advertisement messages are sent periodically but no less than every 3 seconds. To expedite the router discovery procedure, a router advertisement message can be sent as a reply to a router solicitation.

The purpose of DAD is to verify the uniqueness of an IP address prior to assigning it to an interface. As the Neighbor Discovery protocol [20] defines, DAD must be performed on all unicast addresses, regardless of whether they are obtained through stateless autoconfiguration, DHCPv6, or manual configuration. During the DAD process a mobile node sends a number of Neighbor Solicitations (NbSol), each separated by `RetransTimer` milliseconds in order to verify whether any other node has the same address. If after the end of DAD no Neighbor Advertisement message has been received, the mobile node assumes that no other node has this

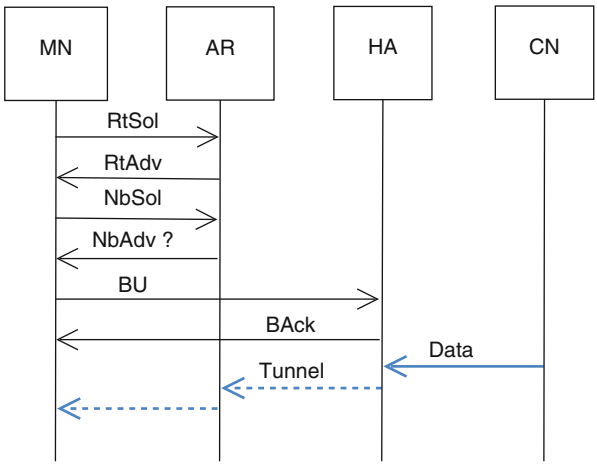


Fig. 9.1 Mobile IPv6 handover signaling

tentative address and assigns it to its interface. The protocol specifies that at least one Neighbor Solicitation must be sent with minimum value of 1,000 ms for the `RetransTimer`. During the DAD procedure the mobile node cannot use the address under test which makes DAD a very time-consuming operation that degrades the handover performance significantly. Alternatively, the optimistic DAD can be used that removes the `RetransTimer` delay during address configuration. The configured address can be used immediately as an Optimistic address which is an address that is assigned to an interface and is available for use, while its uniqueness on a link is being verified. This is based on the assumption that the probability of a conflict is low especially when IEEE identifiers are used.

Once the mobile node moves to a new subnet, it performs configuration of the link-local address, router discovery for new prefix acquisition, and new care-of-address (NCoA) configuration as well as DAD procedure. Subsequently, it registers the NCoA with the home agent through Binding Update (BU) and Acknowledgement (BA) messages. Figure 9.1 depicts the MIPv6 signaling exchange when the mobile node registers its NCoA with the home agent.

9.2.2 Local Mobility Protocols

Although Mobile IP enables the mobile node to maintain its connectivity to the Internet when it roams across IP subnets, it has been slowly deployed in real networks as it suffers from high handover delay. During handover, there is a period in which the mobile node is unable to send or receive packets because of link layer handover latency and IP protocol operations. This handover latency resulting from standard Mobile IP is often unacceptable to real-time traffic. Separating local from global mobility reduces adequately the handover delay. The following extensions are proposed standards of the IETF designed to deal with local mobility and compatible with both versions of Mobile IP protocol.

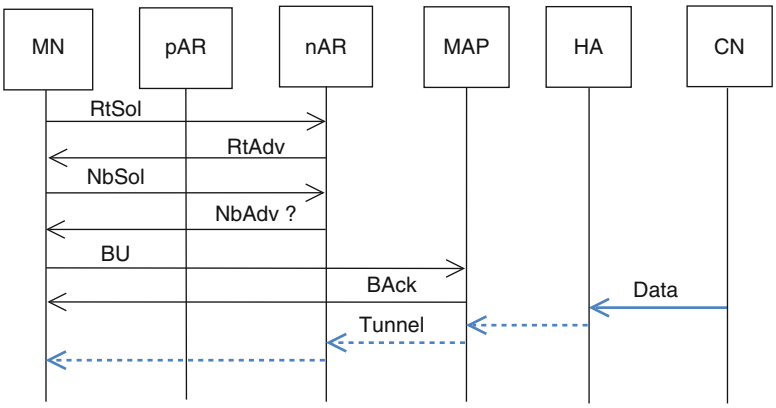


Fig. 9.2 Hierarchical mobile IPv6 handover signaling

9.2.2.1 Hierarchical Mobile IPv6 (HMIPv6)

HMIPv6 [25] introduces hierarchical mobility management in MIPv6, and by that separates local from global mobility. In HMIPv6, global mobility is managed by the MIPv6 protocols, while local handovers are managed locally. To do so, a new Mobile IP functional entity called the mobility anchor point (MAP) is used, which is basically a local HA. As shown in Fig. 9.2, the only difference of HMIPv6 from MIPv6 is that the mobile node registers to the local MAP rather than the HA and, as a result, the registration process is expedited by taking place locally. Figure 9.2 depicts signaling exchange when the mobile node performs a local handover. Consequently, HMIPv6 reduces Mobile IP signaling load and improves the MIPv6 handover delay. All in all, HMIPv6 minimizes the impact on Mobile IP offering additional reliability and scalability.

9.2.2.2 Proxy Mobile IPv6 (PMIPv6)

PMIPv6 [31] is a local mobility protocol which additionally implements mobility management procedures in the network part without involving the mobile node. For that purpose, two new functional entities are introduced in PMIPv6, the local mobility anchor (LMA) and the mobility access gateway (MAG). The LMA is the topological anchor point for the mobile node's home network prefix (i.e. it advertises the prefix to the network), receiving any packets that are sent to the mobile node by any node in or outside the PMIPv6 domain. The LMA is a home agent with enhanced capabilities for supporting PMIPv6. The MAG is a new functional entity that emulates the mobile node's home link on the access link. To do so, the MAG sends router advertisement messages, containing the mobile node's home network prefix. By this means, PMIPv6 does not require the address

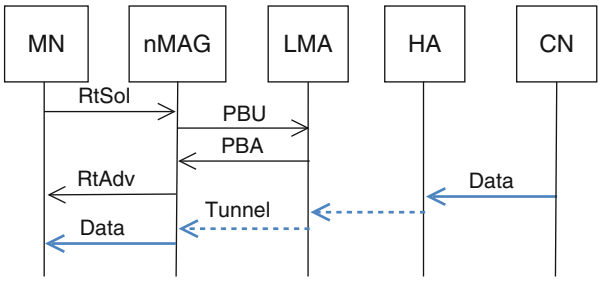


Fig. 9.3 Proxy mobile IPv6 handover signaling

configuration procedure to be performed during handover reducing handover delay and signaling. Typically, MAG is a function runs on an access router. Finally, the advantages of network-based mobility management is that it is compatible with legacy devices and reduces signaling exchange through the wireless link.

Figure 9.3 illustrates the signaling flow for the PMIPv6. When a mobile node enters the serving area of an MAG performs the router discovery procedure by sending a RtSol message. Through this procedure, the MAG obtains the mobile node's identifier such as the link layer identifier and sends a proxy binding update (PBU) message to the LMA. PBU is a BU message extended with a new field to indicate to the LMA that the BU is a proxy registration. New options are available in a PBU message, such as mobile node identifier, access technology type or link-local address. Finally the LMA replies with a proxy binding acknowledgement (PBA) and sets up a bidirectional tunnel with the MAG in order to use the mobile node's home network prefix over it. Upon reception of the PBA, the LMA acquires the mobile node's home network prefix and sends the relevant RtAdv. Receiving the RtAdv, the mobile node assumes that there was no subnet change and the handover process is completed.

9.2.3 Route Optimization

The main goal of MIPv4 was to allow transparent interoperation between mobile nodes and their correspondent nodes using mobility enable agents as described in Sect. 9.2.1.1. Although through this concept the condition of transparency has been satisfied, all datagrams are forced to be routed through the mobile node's home network even if the correspondent node is located at the vicinity of the mobile node. This situation in which correspondent node's packets destined to the mobile node follow a path which is longer than the optimal path because the packets must be forwarded via specific mobility agent is called Triangle Routing [33].

In the case of MIPv4 route optimization is a nonstandard set of extensions [22], while in MIPv6 is a fundamental part of the protocol. However the approach in both cases is similar as the main idea is to allow the mobile node to register its care-of-address not only with the home agent but also with the correspondent node.

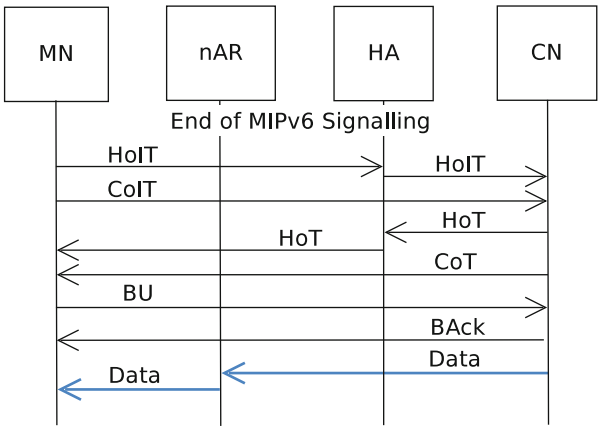


Fig. 9.4 Mobile IPv6 route optimization signaling

As a result a binding cache is maintained by the correspondent node containing the care-of-address of the mobile node in order to optimize the communication path. This is accomplished through the BU and BA messages exchange between the mobile node and the correspondent node.

Nonetheless, before this signaling exchange a mobility security association must be established between the two endpoints of the communication. In MIPv6, this is achieved through another fundamental sub-protocol named Return Routability protocol (RR). The RR is a procedure that enables the correspondent node to carry out a minimal verification that a mobile node owns an address (home address) and is reachable at another (care-of-address). Only with this assurance the correspondent node can accept BU and direct mobile node's data traffic to its claimed care-of-address. The mobile node initiates this testing procedure by sending two messages, one routed through the home agent (Home Test Init—HoTI) and one directly to the correspondent node (Care-of Test Init—CoTI). These messages are used to initiate the RR procedure and convey the home and care-of-address of the mobile node. Upon reception of these two messages the correspondent node produces two keygen tokens based on its secret key and the home or care-of-address. Finally, each token is sent by the correspondent node to the mobile node, one via the home agent (Home Test—HoT) and one directly (Care-of Test—CoT). The RR is completed when the mobile node receives both messages and can now send a BU to the correspondent node. It is noted that, in HMIPv6 described in Sect. 9.2.2.1, the mobile node hides its location (CoA) from both the home agent and correspondent node and, as a result, it cannot use MIPv6 route optimization.

If route optimization is supported by the correspondent node the RR is executed right after the home agent registration and it is followed by the correspondent registration. Figure 9.4 above shows the message flow for the Mobile IPv6 protocol with route optimization.

9.2.4 Fast Handovers

In all previously discussed protocols, a mobile node initiates the network layer handover procedure when it detects that it has moved to a new subnet. Movement detection occurs after link layer handover and as a result link and network handover delays are additive. This can introduce significant disruption which is unacceptable for applications such as 3D video streaming. To address this problem, the IETF Network Working Group has proposed a cross layer solution where link layer triggers are used to optimize the network layer handover procedure. This approach results to a family of protocols defined as fast handovers including specific implementations for MIPv4 [14], MIPv6 [13] and PMIPv6 [31].

A link layer trigger is an event related to the link condition or the link layer handover process. The link layer triggers that are made available to the IP layer are assumed to be generic and technology independent. As these events are very common in cross layer design, IEEE has developed a standard to support handover management and interoperability between heterogeneous network types. The result is the 802.21 MIH protocol, which is a standard that can perform such signaling and assists fast handovers functionality. Such triggers are the "Link Going Down" event for an upcoming change in link layer point of attachment due to signal deterioration, the "Link Down" event which is the trigger that occurs at the previous access router (pAR), informing that the mobile node has moved to another subnet and the "Link Up" is the consequent establishment of the new link.

In the beginning, the mobile node solicits information of the future (new) access router that is going to be handed over. This will result to configuration of a NCoA based on the new access router (nAR) while the mobile node is still connected to its current (previous) access router using its previous care-of-address (PCoA). The next step is to establish a tunnel between the previous and nAR to forward mobile node's data. During this stage buffering can be used to minimize losses. If this establishment occurs while the mobile node is still connected to the pAR the scenario is characterized as the predictive mode of operation. In this case a movement prediction mechanism must be applied. Otherwise, a reactive fast handover occurs meaning that the tunnel starts after the "Link Up" event. Finally, the mobile node registers the NCoA to the home mobility agent starting normal routing of data. However, during this signaling exchange there is no loss of data only an additional delay to forward packets from pAR to the new one.

9.2.4.1 Mobile IPv6 Fast Handovers

In the following, the specific signaling defined for FMIPv6 implementation will be described as shown in Fig. 9.5. Initially, the mobile node sends a Router Solicitation for Proxy Advertisement (RtSolPr) to the pAR to request information for a potential handover. The pAR will respond with a Proxy Router Advertisement (PrRtAdv) providing information about neighboring links and by this way

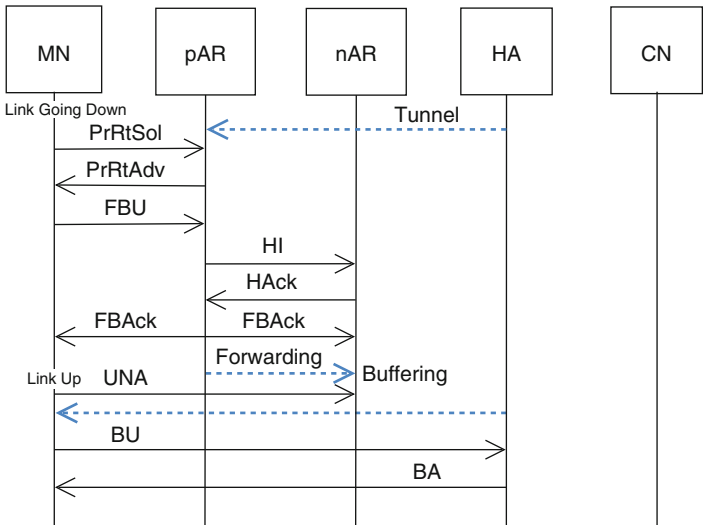


Fig. 9.5 Fast handovers for mobile IPv6 (FMIPv6)

expediting movement detection. After processing the PrRtAdv message, the mobile node sends a fast binding update (FBU) message instructing its pAR to redirect its traffic towards the nAR. The moment of sending the FBU is determined by link-specific events and significantly affects the handover delay. In the predictive mode the mobile node sends the FBU from the pAR's link whenever anticipation of handover is feasible and the pAR responds with a Fast Binding Acknowledgement (FBack) message. A "Link Going Down" event can be such a trigger. From that moment on, packets will be sent to the nAR and will be buffered until the "Link Up" event. Consequently, the smaller the time delay between the "Link Going Down" event and the "Link Down" event the less packets will be buffered and the communication disruption will decrease. In case that anticipation is not feasible or that the mobile node has not received an FBack, the reactive mode is applied where the mobile node sends an FBU immediately after attaching to nARs link.

In any case, the result is that packets arriving for PCoA are tunneled to NCoA. Such tunnel remains active until the mobile node completes the binding update with home agent. In the opposite direction, the mobile node reverse tunnels packets to the pAR in order to ensure that packets containing a PCoA as a source IP address are not dropped due to ingress filtering. Even though the mobile node is IP-capable on the new link, it cannot use the NCoA directly with its correspondents without first establishing a binding cache entry (for the NCoA). It should be noted that route optimization can apply in this case without contributing in the handover delay.

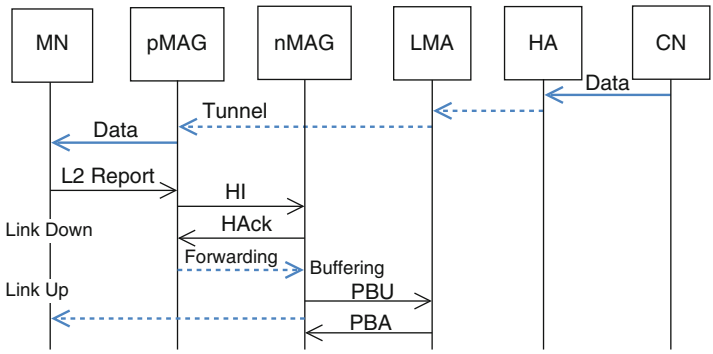


Fig. 9.6 Fast handovers for proxy mobile IPv6 signaling

9.2.4.2 Fast Handovers for Proxy Mobile IPv6 (PFMIPv6)

Fast Handovers can also apply in the case of local mobility management. In PMIPv6 unlike HMIPv6, fast handovers is a proposed standard [31]. In PMIPv6, as described in Sect. 9.2.2.2, a mobile node is not directly involved with IP mobility management operations. Hence, the messages involving the mobile node in FMIPv6 cannot be used in the PMIPv6 context. Therefore, the RtSolPr, the PrRtAdv, FBU, FBack and the UNA messages are not applicable and they are omitted.

On the other hand, the protocol operations are transparent to the LMA. Note that the mobile node is capable of reporting lower-layer information to the nAR meaning that MIH can be applied. Moreover, when the mobile node establishes a new link layer connection its identifier is also transferred to the new MAG (nMAG). This can be regarded as a substitute for the UNA message. The sequence of message exchanges for the predictive fast handover is illustrated in Fig. 9.6.

9.3 MIH and Network Coding

Before presenting our approach which integrates network coding into MIH protocol, the main characteristics of MIH protocol are shown and some concepts of network coding are described in the next subsections.

9.3.1 Media Independent Handover Protocol

According to IEEE 802.21, MNs and network entities communicate with each other using MIH protocol messages. MIH protocol is employed to remotely send messages between separate media independent handover function (MIHF) entities.

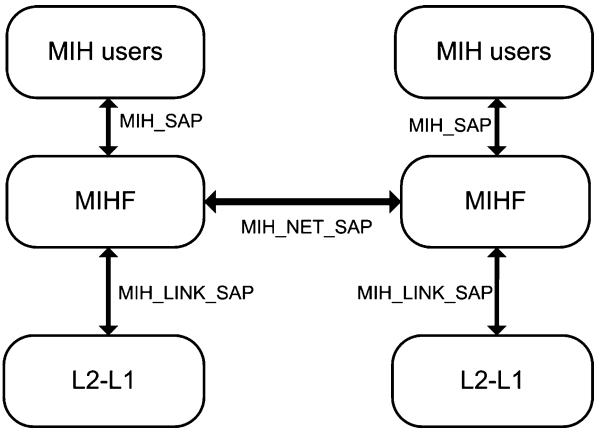


Fig. 9.7 IEEE 802.21 reference model with the different kind of interfaces used to communicate with different layers and entities

Figure 9.7 represents the IEEE 802.21 reference model and types of MIHF relationship. Once two MIHF entities need to communicate with each other, a transaction is initialized. An MIH transaction is a flow of messages with the same Transaction-ID submitted to, or received from, a particular MIHF ID. A specific MIH node cannot have more than one transaction pending for each direction with an MIH peer. According to classical MIH protocol, if the remote communication between entities is not reliable, an acknowledgement service is required. The MIH acknowledgement service uses two bits inside the MIH header: the ACK-Req bit is set by the source and the ACK-Rsp bit is set by destination. After sending an MIH protocol message with ACK-Req bit set, the source starts a retransmission timer and keeps a copy of it while the timer is active. If the acknowledgement message is not received before the expiration of the timer, the source node retransmits the saved frame with the same Message-ID, with the same Transaction-ID and with ACK-Req bit set. Otherwise, when acknowledgement is received before expiration of the timer or before any other retransmission attempt, the source ensures the correct reception of the message, resets the timer, and deletes the saved copy of the packet. Moreover, if the MIH source receives ACK for any of the previous transmission attempts, then the communication is classified as successful and it does not need to wait for any further acknowledgements. Retransmissions are done while ACK is received or the number of retransmissions reaches the maximum value.

On the destination side, frames received with the ACK-Req bit set cause the return of an MIH acknowledgement message with ACK-Rsp bit set in the header and with the same Message-ID and Transaction-ID. In particular, acknowledgements are MIH packets without payload. When an immediate response is waited by the MIH source, the receiver sends the corresponding MIH answer message with ACK-Rsp bit set. Then, the destination can also set the bit ACK-Req to ask the source to acknowledge the response message. If multiple messages are received, the sink only processes the first one. In this sense, all the duplicate frames are acknowledged. All MIH protocol messages have two identifiers: MIHF ID

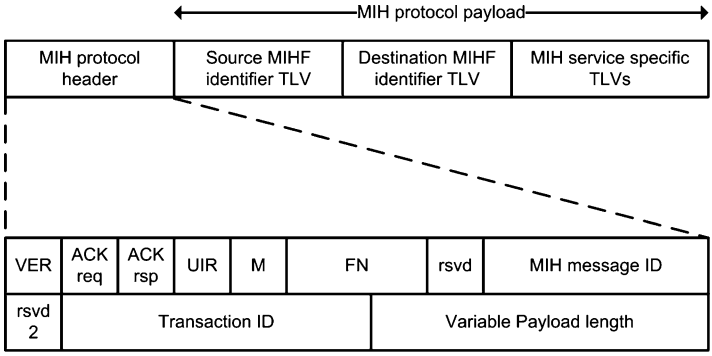


Fig. 9.8 MIH general protocol frame format and its header [8]

Table 9.1 MIH protocol header fields [8]

Field name	Size[a]	Description
Version	4	It specifies the current version of MIH protocol in use.
ACK-Req	1	It is used to request an acknowledgment for the message.
ACK-Rsp	1	It is used to answer to the request for and acknowledgment message.
Unauthenticated Information request (UIR)	1	It informs the MIH Information Service if the message is sent in pre-authentication/pre-association state so that the length of the response message is limited.
More fragment (M)	1	It is used if the message is a fragment to be followed by another fragment.
Fragment number (FN)	7	It is used to represent the sequence number of the fragment (0–127).
Reserved (rsvd)	1	It is kept reserved and usually set to "0".
MIH message ID	16	It is a combination of three different fields: Service identifier (SID), operation code (Opcode) and action identifier (AID). The first identifies the different MIH services. The second is the kind of operation to be performed with respect to the SID. The third indicates the action to be done according to the SID.
Reserved2 (rsvd2)	4	It is kept reserved and all the bits are usually set to "0".
Transaction-ID	12	It is used to match requests, responses and acknowledgments.
Variable payload length	16	It defines the total length of the variable payload of the respective MIH protocol frame.

[a] Sizeis expressed in bits

and Transaction-ID. The former uniquely identifies an MIHF entity to provide the services whilst the latter matches a request message with the correspondent response message or acknowledgement. An MIH protocol payload is constituted by a Source MIHF Identifier Type-Length-Value (TLV), a Destination MIHF Identifier TLV and an MIH Service Specific TLVs. Figure 9.8 depicts the fields of MIH protocol frame format and of its header. Table 9.1 describes the meaning of the different fields of the MIH protocol header format.

9.3.2 Network Coding

Network coding can represent an efficient solution to make MIH more reliable. The idea of coding information at network layer started in 2000 with [1]. This work showed that higher rates could be achieved in a network by mixing packets together at the nodes instead of only forwarding them. Next, [16] demonstrated that linear combination of packets were optimum solution to design network codes for multicast scenarios. These initial approaches were mainly taking advantage of graph theory and linear algebra. A complete algebraic framework for network coding was proposed by [12]: its theoretic results were fundamentals for the development of random linear network coding (RLNC) [7].

Besides improving throughput, network coding can be useful for error correction. In particular, [5, 30] defined the capabilities and theoretic characteristics of NEC codes. Then, subspace coding [11, 24] emerged as a new way of designing NEC codes, without the need of conveying the coefficients of the linear combinations into the header of the packets. This noncoherent approach communicates information via subspaces and exploits the fact that network linear operations do not affect them.

Despite the amount of achievements in network coding theory, the research on network coding for practical application is still at the beginning: it is not yet clear how network coding can be developed into practice and the effects it can produce on existing systems. The first work focused on practical network coding was [6]. Next, the landmark articles [26, 27] proposed a mechanism to efficiently incorporate network coding into the TCP. The novel TCP/NC consists in a modified protocol stack and in an innovative acknowledgement system, which makes TCP compatible with coding operations on packets. Side by side, [3, 15, 17, 29] investigated how retransmission systems, such as automatic repeat request (ARQ) and hybrid automatic repeat request (HARQ), could meet network coding. The two main solutions to modify the acknowledgement mechanisms were:

- once a packet is not received, the source retransmits the linear combination of that packet with a new one,
- if a set of packets is received with errors, the source retransmits their linear combination.

Next, [10] described a cooperative radio access network (RAN) MAC layer protocol based on RLNC to ensure reliable and flexible data delivery over 3GPP long-term evolution-advanced (LTE-A) RAN.

Section 9.3.4 describes a novel NC-MIH protocol based on subspace coding. The main idea is to deploy network coding to avoid erasures and to reduce retransmission load, acknowledgments, and energy consumption. The network coding operations are efficiently integrated into the original MIH protocol and a novel acknowledgment mechanism is shown.

9.3.3 Subspace Coding

Coding through subspaces [11] represents an alternative solution to RLNC to develop NEC codes independent from network topology. In particular, subspace coding in respect of RLNC, does not require appending coefficients of linear combinations to headers of packets. This is important to avoid additional overhead in transmissions, especially in wireless scenarios.

In order to explain the main characteristics of subspace codes, let us define the channel model

$$\mathbf{Y} = \mathbf{HP} + \mathbf{GE} \tag{9.1}$$

where the rows of matrices **P** and **E** are, respectively, the source packets and error packets, **H** and **G** are random matrices and **Y** are the received packets. Let W be a fixed N-dimensional vector space over a finite field of size q. An operator channel associated with the ambient space W is a channel input and output alphabet $\mathscr{P}(W)$. Given as channel input the subspace V and as channel output the subspace U, the relationship between them is

$$U = \mathscr{H}_k(V) \oplus E \tag{9.2}$$

where $k = \dim(U \cap V)$ and E is an error space. Especially, the operator channel transforms V into U by committing $\rho = \dim(V) - k$ erasures and $t = \dim(E)$ errors. Next, it is fundamental to define a metric function for subspace coding. In order to do that, let d be a function $d : \mathscr{P}(W) \times \mathscr{P}(W) \rightarrow Z_+$, with Z_+ the set of nonnegative integers. Then, the function

$$d(A, B) := \dim(A + B) - \dim(A \cap B) \tag{9.3}$$

is the distance between the two subspaces A and B. Finally, a code C for an operator channel with ambient space W is a nonempty collection of subspaces of W. Other two important parameters of subspace codes are the maximum distance between codewords (subspaces)

$$L(\mathrm{C}) := \max_{A \in C} \dim(A) \tag{9.4}$$

and the rate of the code

$$R = \frac{\log_q(|\mathrm{C}|)}{NL} \tag{9.5}$$

where NL denotes either the number of q-ary symbols contained in the source packets or the number of required channels to inject L source packets.

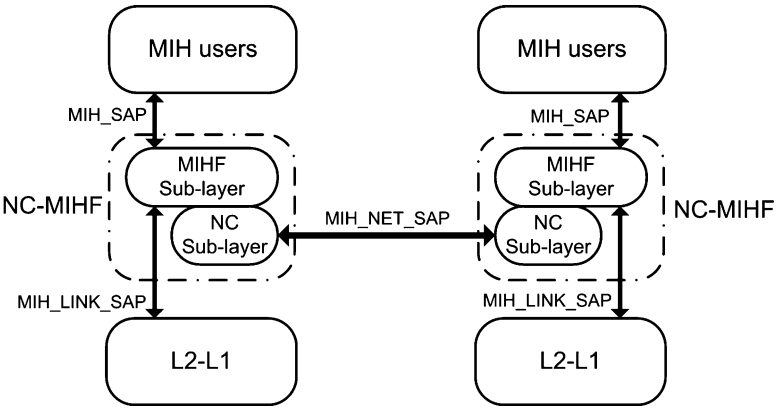


Fig. 9.9 NC-MIH reference model showing the two new MIHF and NC sub-layers and their interfaces

The first algorithm to efficiently construct subspace codes was provided by [11] in 2008. Because of the name of the authors, these codes have been called KK codes. The basis for the transmitted vector space is obtained via the evaluation of a linearized message polynomial in a Reed-Solomon-like way. On the other side, KK codes decode received subspaces via a Sudan-style algorithm [19].

9.3.4 NC-MIH Protocol

This section presents our approach to merge subspace coding with the original MIH protocol. The reason to incorporate network coding into MIHF layer is the reduction of packet erasures due to unreliable links and the reduction of losses caused by vertical handovers. The existing protocol stack of MIHF is modified to efficiently embed coding operations. So, the classical MIHF layer defined by [8] is split into two sub-layers: an MIHF sub-layer and a network coding (NC) sub-layer. The former has kept almost all the functionalities that belonged to the previous MIHF layer: in fact, this layer directly communicates with upper and lower layers, respectively, through the interfaces MIH_SAP and MIH_LINK_SAP. The NC sub-layer is placed under MIHF sub-layer to perform encoding and decoding operations over the MIH frames. Because of this, MIH_NET_SAP interface has been moved to NC sub-layer to allow this sub-layer to exchange information between remote MIHF entities. Figure 9.9 depicts how the reference model in Fig. 9.7 has been modified. At the source, the MIH sub-layer sends a set of packets—belonging to the same transaction—to the NC sub-layer which stores them in an encoding buffer. Then, the KK encoder encodes the buffered packets according to a fixed subspace codebook: the dimensionality of the encoded subspaces cannot exceed the min-cut bound of the connection.

When subspaces are sent, a copy of them is stored and a retransmission timer starts. This procedure allows NC sub-layer to perform retransmissions in case of errors and erasures. It is useful to describe the coding procedure at the MIH source entity. First, L linearly independent vectors $\mathbf{a}_1, \ldots, \mathbf{a}_L$ are given: these vectors span an L-dimensional subspace over a finite field of size q. Source frames, which are created at MIH sub-layer, are represented by the matrix $\mathbf{U}$ with k rows

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_0 \\ \vdots \\ \mathbf{p}_{k-1} \end{bmatrix} \tag{9.6}$$

where the elements are belonging to a finite field of size q^m. Based on these k vectors that are collected in the buffer, the encoder forms a linearized polynomial

$$f(\mathbf{x}) = \sum_{i=0}^{k-1} \mathbf{p}_i \mathbf{x}^{q^i}. \tag{9.7}$$

Then, the encoder evaluates the function $f(\mathbf{x})$, with $\mathbf{x} = \mathbf{a}_1, \ldots, \mathbf{a}_L$, and calculates the set $\mathbf{b}_1, \ldots, \mathbf{b}_L$, with $\mathbf{b}_i = f(\mathbf{a}_i)$. Since vectors $\mathbf{a}_1, \ldots, \mathbf{a}_L$ are linearly independent, the tuples $(\mathbf{a}_1, \mathbf{b}_1), \ldots, (\mathbf{a}_L, \mathbf{b}_L)$ are linearly independent as well and they result in a basis of a vector space V of dimension $L + m$ over a finite field of size q. Finally, subspace V is sent via the MIH_NET_SAP interface. It is important to underline that subspaces, obtained by $f(\mathbf{x})$ for different set of source messages, are all distinctive if the condition $\mid A \mid \geq k$ is satisfied.

Once the destination receives the packets, the NC sub-layer decodes the subspaces by comparing them with the ones listed in the codebook and sends the information packets to the MIH sub-layer. Side by side, the decoder also acknowledges the subspaces that are correctly received. The ACKs are related to each subspace and are interpreted by the NC sub-layer at the source. In fact, from the moment an acknowledgment[1] is received, the NC sub-layer releases the encoded packet stored before and stops the respective retransmission timer. Moreover, it removes the information packets which span that subspace from the buffer to give space to new packets to be encoded and sent. The MIH destination entity receives a subspace U of dimension $r = L - \rho + t$, where the tuples $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, r$, are a basis for U. First, the decoder constructs a nonzero bivariate polynomial

$$Q(\mathbf{x}, \mathbf{y}) = Q_x(\mathbf{x}) + Q_y(\mathbf{y}) \tag{9.8}$$

where $Q(\mathbf{x}_i, \mathbf{y}_i) = 0, 1 \leq i \leq r$. Two assumptions are important to find the solution of a homogeneous system composed of r equations in $2\tau - k + 1$ variables: $Q_x(\mathbf{x})$ is

[1] Acknowledgments are related to a unique subspace.

a linearized polynomial over finite field of size q^m of degree at most $q^{\tau-1}$ and $Q_y(\mathbf{y})$ is a linearized polynomial over finite field of size q^m of degree at most $q^{\tau-k}$. Because of that, it is achieved a nonzero solution if $r = L - \rho + t < 2\tau - k + 1$. Another condition is that $Q(\mathbf{x}, f(\mathbf{x})) = 0$. It is possible to convert (9.8) into

$$Q_y(\mathbf{x}) \otimes f(\mathbf{x}) + Q_x(\mathbf{x}) = 0. \tag{9.9}$$

From this relation, polynomial $f(\mathbf{x})$ is calculated by using Euclidean algorithm. If no polynomial is found, the decoding procedure fails. Finally the output $f(\mathbf{x})$ corresponds to the codeword $\hat{V} \in \mathrm{C}$ if $d(U, \hat{V}) < L - k + 1$. The time complexity of the decoding procedure is $O((L + m)^2)$ operations over a finite field of size q^m. Once the codeword $\hat{V}$ is recovered, the NC sub-layer at the receiver sends back an acknowledgement referred to subspace V. Side by side, the decoder interprets—via the codebook—the source packets from $\hat{V}$ and finally, the packets are ready for the MIH sub-layer.

MIH protocol leaves reliability issues for the particular transport layer protocols implemented. When the message loss rate experienced on the link is greater than or equal to 0. 01 %, the acknowledgement service is deployed to make the communication reliable. The main advantage of our NC-MIH protocol is that subspace coding within the MIHF layer guarantees the reliability of the protocol, independently from the transport protocol in use. The integration of a coding sub-layer to provide less errors becomes more important when, for example, user datagram protocol (UDP) is the transport protocol. In case, the system required UDP to reduce latencies for delay-sensitive handover communications—while reducing reliability—an NC-MIH sub-layer could achieve low losses together with low latencies.

9.3.5 Practical Implementation

After previous theoretic discussion about how to integrate subspace coding into MIH protocol, this section describes our approach to implement in practice the NC-MIH protocol.

First, since the one-to-one correspondence between byte and symbol is assumed, coding operations are performed over a finite field of size 256. Next, the information messages considered in Sect. 9.3.4 are of fixed length. However, MIH protocol defines a variable payload length. Therefore, while the encoder takes the source packets from the buffer, it introduces zero symbols to achieve a fixed size packet. The reference to perform this action is the longest packet in the encoder buffer. Nevertheless, the variable payload length field inside packet headers remains the same to inform the receiver about the real packet size of each frame.

An encoded packet is constituted by either packets in the buffer or a subset of them. Its header has the aim of helping its routing towards the destination and to help the receiver to manage the classification of the packet and the acknowledgement process. Figure 9.10 depicts the structure of an encoded packet.

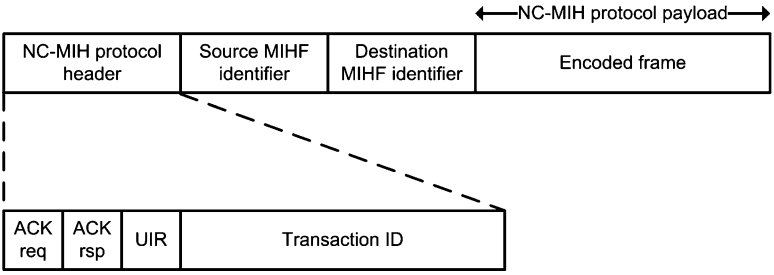


Fig. 9.10 NC-MIH protocol frame format

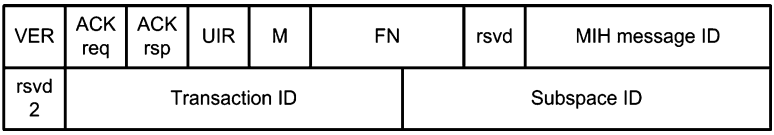


Fig. 9.11 NC-MIH acknowledgement frame format

- *Ack-Req and Ack-Rsp*: they allow the receiver to perform acknowledgement mechanisms of encoded packets, once received.
- *Unauthenticated information request*: it shows the receiver if the packets encoded are from an MIHF entity that is communicating in unauthenticated mode.
- *Transaction-ID*: it indicates the transaction the encoded message belongs to.
- *Source and Destination MIHF identifier*: they uniquely identify source and destination of the message.

The information packets stored in the source buffer have associated with a retransmission timer. When an ACK is interpreted by the NC sub-layer, the packets which span that subspace are removed from the buffer and their retransmission timers are cancelled. After a successful decoding process, the NC sub-layer at the receiver acknowledges the subspace received. Figure 9.11 shows the structure of the acknowledgement frame. As IEEE 802.21 standard states, the ACK frame is only constituted by MIH frame header. In case of NC-MIH protocol, a field of 16 bits called Subspace-ID is introduced to create a single correspondence between subspaces and acknowledgements. In fact, the Subspace-ID field indicates the number of the subspace in the codebook list and uniquely identifies it.

It is possible to have intermediate nodes between source and destination. Once encoded packets are received, they are linearly combined together in an arbitrary way. Since linear operations are not affecting subspaces, the receiver is still able to decode the original message.

9.3.6 *Performance Analysis – An Energy Perspective*

This section presents a first analysis to show the differences in energy consumption between classical MIH protocol and our NC-MIH protocol. The simple topology considered is a point-to-point communication network with a source MIH entity sending packets to a destination MIH entity. The model of the link consists in an erasure channel, in which each packet is lost with a constant probability ε in transit independent of all the other packets. Then, the capacity of the erasure channel is $1 - \varepsilon$ and the range of possible code rates achievable becomes $R < 1 - \varepsilon$. This scenario can be seen as a MN communicating with a point-of-attachment (in this case an IEEE 802.11n access point). When MIH protocol is implemented, the consumption calculated considers the presence of acknowledgment service or the application of a link-layer efficient erasure code [34] based on cascades of sparse bipartite graphs.

Let us assume the packet erasure probability of the wireless link $\varepsilon = 0.1$ and the real rate of the link $r_{\max} = 80$ Mb/s. The MIH source entity has to send 7, 000 packets of constant size 256 bytes to the MIH destination entity. Table 9.2 summarizes the energy consumption per second of IEEE 802.11n for each state. As Sect. 9.3.1. has shown above, the size of an ACK frame is 8 bytes. The following evaluation does not take into account latencies and retransmissions caused by the expiration of retransmission timers.

First, the system deploys MIH protocol with acknowledgment service to make the transmission reliable. Each packet transmitted requires an acknowledgment and the frames lost need to be retransmitted and acknowledged. The second case is the one in which there are no losses because of the packet erasure code [34]. A code of at least rate $\frac{9}{10}$ guarantees to recover all the lost packets after the decoding procedure. However, this produces a redundancy of 778 frames plus relative ACKs. The last case investigates the use of NC-MIH protocol instead of classical MIH protocol. The extra overhead into NC-MIH frames includes 15 bits per frame. Figure 9.12 depicts the energy consumption per bit of the three previous approaches by using the energy consumptions summarized in Table 9.2. It is possible to see that NC-MIH protocol has the least energy consumption. It is important to highlight that, in case of MIH protocol with erasure code and NC-MIH protocol, the energy consumption calculated represents an upper bound: in fact, the analysis considered the application of the acknowledgment service. Nevertheless, if the communication is reliable because of coding procedures, acknowledgments become unnecessary.

Table 9.2 Energy consumption measurements of IEEE 802.11n by considering unit time 1*s* [23]

Device state	Energy consumption [mJ]
Off	0
Receive	1, 270
Transmit	1, 990

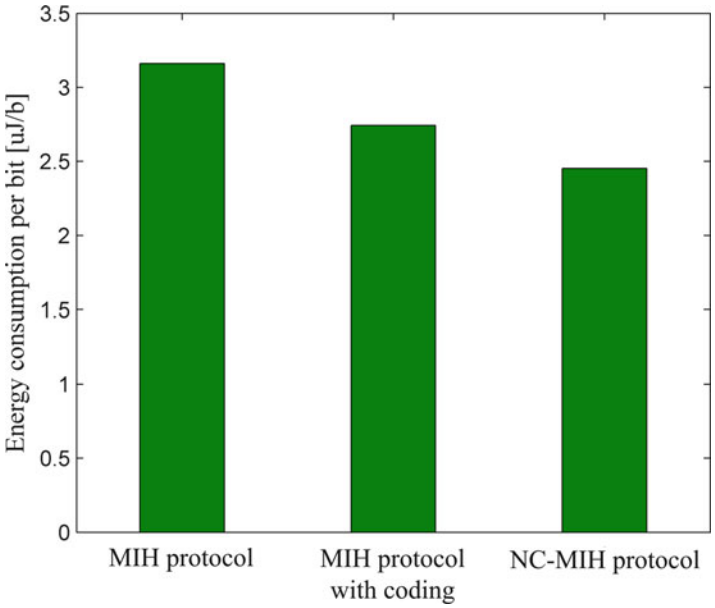


Fig. 9.12 Energy consumption per bit of the systems which use MIH protocol, MIH protocol with a packet erasure code and NC-MIH protocol

9.4 IP Mobility Protocols Analysis

In the following section the aforementioned mobility management protocols will be analyzed in terms of Signaling Cost, Packet Delivery Cost, and Handover Disruption Time. To perform the analysis, the fluid flow mobility model will be adopted. Fluid flow model is appropriate for mobility nodes with high mobility. The model is applicable under the following assumptions:

- Subnet and domain areas are circular.
- The density of mobile nodes is uniform in these areas.
- The direction of mobile node's movement is uniformly distributed in the range of $(0, 2\pi)$.

Let v be the nodes' average speed, R is the coverage radius of each wireless access router and N is the number of subnets in a domain. Then the cell crossing rate μ_c, the inter-domain crossing rate μ_d and the intra-domain crossing rate μ_s are expressed as follows [2].

$$\mu_c = \frac{2 \cdot v}{\pi \cdot R}\ ,\ \mu_d = \frac{\mu_c}{\sqrt{N}} \text{ and } \mu_s = \mu_c \cdot \frac{\sqrt{N} - 1}{\sqrt{N}} \tag{9.10}$$

Table 9.3 Length of control messages (in bytes)

Message	Length	Message	Length	Message	Length	Message	Length
AgSol	28	HoTI	64	BA (CN)	66	PBA	76
AgAdv	67	CoTI	64	BU (MAP)	56	HI	52
RReq	60	HoT	74	BA (MAP)	56	HAck	52
RRpl	56	CoT	74	FBU	56	PrRtSol	52
IPv6	40	BU (HA)	56	FBA	56	PrRtAdv	80
RtSol	52	BA (HA)	56	UNA	52	NbSol	52
RtAdv	80	BU (CN)	66	PBU	76	data	1,460

Assuming that the session arrival process follows a Poisson distribution with parameter λ_s and from (9.10), the average number of movements for each type of previously discussed crossing during an inter-session arrival can be obtained.

$$E(N_c) = \frac{\mu_c}{\lambda_s}\ ,\ E(N_d) = \frac{\mu_d}{\lambda_s} \text{ and } E(N_s) = \frac{\mu_s}{\lambda_s} \tag{9.11}$$

Finally, for our analysis the session to mobility ratio (SMR) will be utilized. The SMR is defined as the ratio of the session arrival rate to the mobility rate which is the inverse of $E(N_c)$. If the SMR is high, the session activity is a more important factor than the mobility rate and hence reduction of the packet delivery cost is more preferable rather than the signaling cost.

9.4.1 Signaling Cost Analysis

The transmission cost of control message M between nodes X and Y can be expressed as:

$$C^M_{X,Y} = a_{\text{wl}} \cdot L_p + a_{\text{w}} \cdot L_p \cdot (d_{X,Y} - 1). \tag{9.12}$$

where a_{wl} and a_{w} is the transmission unit cost in the wireless and wired link, respectively, L_p is the message size in bits and $d_{X,\ Y}$ the number of hops between nodes X and Y. In this analysis, the message processing cost in routers is considered negligible. The size of messages is given in Table 9.3 whereas all the network related parameters are given in Table 9.4. As described in Sect. 9.2.2 HMIPv6 and PMIPv6 perform local binding updates in intra-domain crossing. For inter-domain crossings although not explicitly defined in the current specifications, it is assumed that the MIPv6 mobility service is used. Thus, the average signaling cost for local mobility protocols between two consecutive session arrivals is given by:

$$C_{\text{overall}} = E(N_s) \cdot C_{\text{local}} + E(N_d) \cdot C_{\text{global}} \tag{9.13}$$

Table 9.4 Network related parameters

Parameter	Value (hops)	Parameter	Value	Parameter	Value
$d_{\mathrm{HA,AR}}$	6	B_{wl}	11 Mbps	q	0.01
$d_{\mathrm{HA,CN}}$	4	B_{w}	100 Mbps	$\bar{\omega}_q$	0.1 ms
$d_{\mathrm{CN,AR}}$	9	L_{wl}	10 ms	v	10 m/s
$d_{\mathrm{HA,GW}}$	4	L_{w}	2 ms	N	30 subnets
$d_{\mathrm{GW,AR}}$	3	a_{wl}	0.0015	R	150 m
$d_{\mathrm{pAR,nAR}}$	2	a_{w}	0.001	D_{L2}	100 ms

Obviously, global mobility protocols perform the same process regardless the type of crossing. The average signaling cost in this case is:

$$C_{\mathrm{overall}} = E(N_c) \cdot C_{\mathrm{global}} \tag{9.14}$$

Subsequently, the signaling cost for MIPv6 using the optimistic DAD concept and without performing the return routability process is:

$$\begin{aligned} C^{\mathrm{MIPv6}} &= E(N_c) \cdot \left(C^{\mathrm{RtSol}}_{\mathrm{MN,AR}} + C^{\mathrm{RtAdv}}_{\mathrm{AR,MN}} + C^{\mathrm{NbSol}}_{\mathrm{MN,AR}} + C^{\mathrm{BU}}_{\mathrm{MN,HA}} + C^{\mathrm{BA}}_{\mathrm{HA,MN}} \right) \\ &= E(N_c) \cdot C^{\mathrm{MIPv6}}_{\mathrm{single}}. \end{aligned} \tag{9.15}$$

If the route optimization procedure is used, then the signaling cost is:

$$\begin{aligned} C^{\mathrm{MIPv6_RO}} = E(N_c) \cdot \Big(& C^{\mathrm{MIPv6}}_{\mathrm{single}} + C^{\mathrm{HoTI}}_{\mathrm{MN,HA}} + C^{\mathrm{HoTI}}_{\mathrm{HA,CN}} + C^{\mathrm{CoTI}}_{\mathrm{MN,CN}} \\ & + C^{\mathrm{HoT}}_{\mathrm{CN,HA}} + C^{\mathrm{HoT}}_{\mathrm{HA,MN}} + C^{\mathrm{CoT}}_{\mathrm{CN,MN}} + C^{\mathrm{BU}}_{\mathrm{MN,CN}} + C^{\mathrm{BA}}_{\mathrm{CN,MN}} \Big). \end{aligned} \tag{9.16}$$

Based on the previous analysis the signaling cost for HMIPv6 is:

$$\begin{aligned} C^{\mathrm{HMIPv6}} = E(N_s) \cdot \Big(& C^{\mathrm{RtSol}}_{\mathrm{MN,AR}} + C^{\mathrm{RtAdv}}_{\mathrm{AR,MN}} + C^{\mathrm{NbSol}}_{\mathrm{MN,AR}} + C^{\mathrm{BU}}_{\mathrm{MN,MAP}} \\ & + C^{\mathrm{BA}}_{\mathrm{MAP,MN}} \Big) + E(N_d) \cdot C^{\mathrm{MIPv6}}_{\mathrm{single}}. \end{aligned} \tag{9.17}$$

In the case of FMIPv6, it is worth noting that DAD is not performed as the AR has a pool of addresses which defends for this reason. So the signaling cost is:

$$\begin{aligned} C^{\mathrm{FMIPv6}} = E(N_c) \cdot \Big(& C^{\mathrm{PrRtSol}}_{\mathrm{MN,pAR}} + C^{\mathrm{PrRtAdv}}_{\mathrm{pAR,MN}} + C^{\mathrm{FBU}}_{\mathrm{MN,pAR}} \\ & + C^{\mathrm{HI}}_{\mathrm{pAR,nAR}} + C^{\mathrm{HAck}}_{\mathrm{nAR,pAR}} + C^{\mathrm{FBA}}_{\mathrm{pAR,MN}} + C^{\mathrm{FBA}}_{\mathrm{pAR,nAR}} + C^{\mathrm{UNA}}_{\mathrm{MN,nAR}} \\ & + C^{\mathrm{BU}}_{\mathrm{MN,HA}} + C^{\mathrm{BA}}_{\mathrm{HA,MN}} \Big). \end{aligned} \tag{9.18}$$

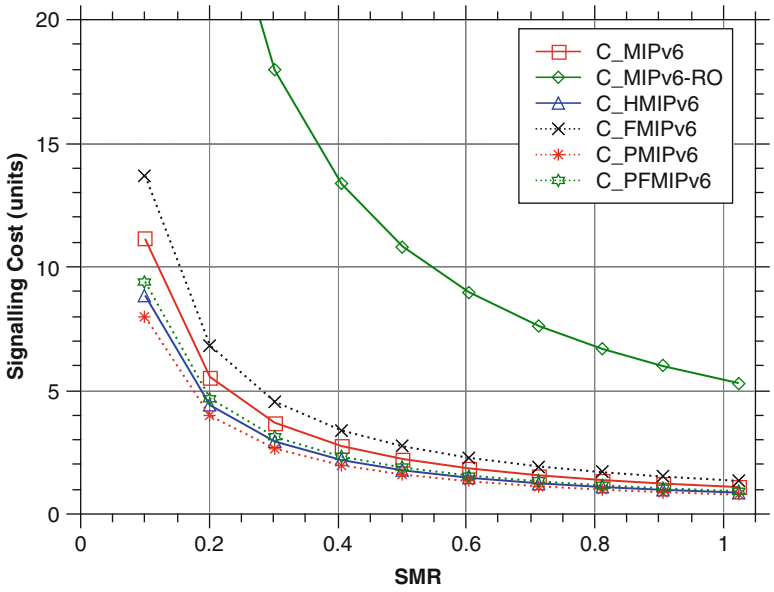


Fig. 9.13 Impact of SMR on handover signaling cost

In the case of PMIPv6 the signaling is significantly reduced to:

$$C^{\text{PMIPv6}} = E(N_s) \cdot \left(C^{\text{RtSol}}_{\text{MN,MAG}} + C^{\text{PBU}}_{\text{MAG,LMA}} + C^{\text{PBA}}_{\text{LMA,MAG}} + C^{\text{RtAdv}}_{\text{MAG,MN}} \right) + E(N_d) \cdot C^{\text{MIPv6}}_{\text{single}}. \tag{9.19}$$

Finally, regarding the PFMIPv6 protocol:

$$C^{\text{PFMIPv6}} = E(N_s) \cdot \left(C^{\text{HI}}_{\text{pMAG,nMAG}} + C^{\text{HAck}}_{\text{nMAG,pMAG}} + C^{\text{PBU}}_{\text{nMAG,LMA}} + C^{\text{PBA}}_{\text{LMA,nMAG}} + C^{\text{RtSol}}_{\text{MN,nMAG}} + C^{\text{RtAdv}}_{\text{nMAG,MN}} \right) + E(N_d) \cdot C^{\text{MIPv6}}_{\text{single}}. \tag{9.20}$$

Figure 9.13 illustrates the binding update signaling cost during handover as a function of SMR based on the previous analysis. When SMR is less than the unit, the cell crossing rate is greater than session arrival rate. Hence, more handovers are performed between sessions and the signaling overhead increases. As the mobility rate decreases (i.e. average number of handovers) the cost of signaling declines too. The highest signaling cost is incurred by route optimization in MIPv6 due to the extensive message exchange imposed by the return routability process. However, as it will be discussed later in Sect. 9.4.2, in route optimization the packet delivery cost decreases as well as the packet end-to-end delay. This is the trade-off for enabling direct communication between mobile and correspondent node. It is noted

that packet end-to-end delay is a crucial aspect regarding time sensitive applications such as 3D video streaming. On the other hand, local mobility protocols achieve the lowest costs as the handover signaling during intra-domain crossings takes place locally. In particular, PMIPv6 performs better than HMIPv6 as no messages are transmitted over the air due to its network-based approach. Finally, the fast handovers schemes increase the signaling cost compared to the basic protocols due to messages introduced for handover anticipation. As it will be seen later, this overhead of fast handover schemes is traded off by lower handover delay and packet loss. Once again, local mobility (i.e. PFMIPv6) performs better than global (i.e. FMIPv6). In particular, PFMIPv6 manages mobility locally and without evolving the mobile node, thus becoming a very attractive solution.

9.4.2 Packet Delivery Cost

According to packet delivery cost, mobility protocols can be classified into three scenarios based on the path that is used to send packets to the mobile node. The first case contains all the global mobility protocols with no route optimization support. Therefore all traffic is intercepted by the home agent and then tunneled to the communication endpoint. On the contrary, when route optimization is in place, traffic bypasses the home agent and data follows the direct path between the mobile node and its correspondent node. Lastly, there is the case of local mobility where route optimization cannot be applied due to the fact that the local binding hides the global routable address of the mobile node. Accordingly, data traffic at first is directed to the home agent and then is forwarded to the local mobility agent before being delivered to the node. Following the previous cost analysis and assuming that the average number of packets per session is $E(P)$, the packet delivery cost for global mobility protocols without route optimization is:

$$\mathrm{PC}^{\mathrm{Global}} = E(P) \cdot L_p \cdot \left(C^{\mathrm{data}}_{\mathrm{CN,HA}} + C^{\mathrm{encap}}_{\mathrm{HA,MN}} \right). \tag{9.21}$$

When route optimization is applied:

$$\mathrm{PC}^{\mathrm{RO}} = E(P) \cdot L_p \cdot C^{\mathrm{data}}_{\mathrm{CN,MN}}. \tag{9.22}$$

In the end, regarding local mobility protocols:

$$\mathrm{PC}^{\mathrm{Local}} = E(P) \cdot L_p \cdot \left(C^{\mathrm{data}}_{\mathrm{CN,HA}} + C^{\mathrm{encap}}_{\mathrm{HA,GW}} + C^{\mathrm{encap}}_{\mathrm{GW,MN}} \right). \tag{9.23}$$

Figure 9.14 depicts the packet delivery cost as function of the average size of the session in packets. While in terms of signaling cost the local mobility accomplishes better results, in terms of packet delivery cost and the consequent packet end-to-end delay performs worst. This is because packets routed between the home agent and

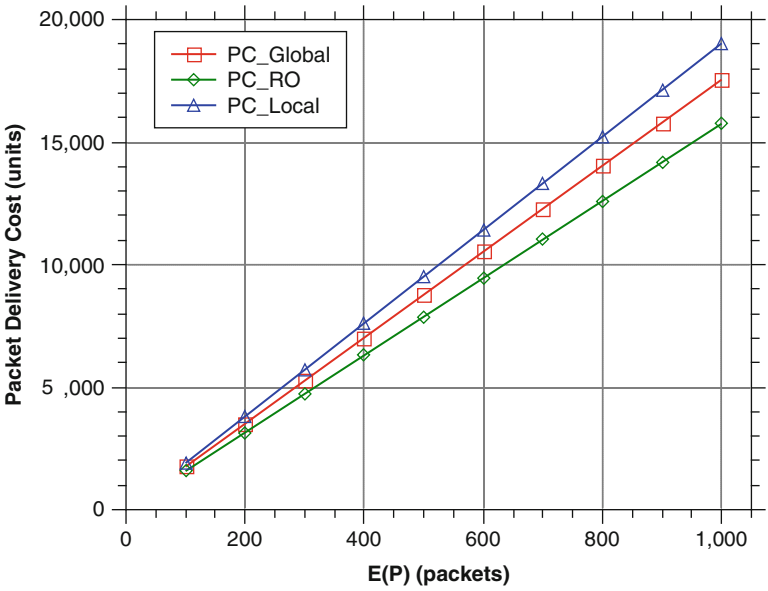


Fig. 9.14 Impact of session size on packet delivery cost

the mobile node are intercepted by the local gateway (GW) instead of following the optimal direct path. However, as the distance between the home agent and the local gateway increases the impact of this triangular routing decreases. Conversely, route optimization is the most efficient solution with respect to packet delivery cost although the worst in regard to signaling cost.

9.4.3 Handover Delay

The handover delay is defined as the time interval between the last packet received from the pAR and the first packet received through the new link. The handover delay comprises the link layer handover delay D_{L2}, network layer handover delay and the time needed to transmit a packet to the mobile node through the new point of attachment. As explained in Sect. 9.2.4, these delays are additive with the exception of fast handovers scenarios where the network layer handover operations take place before the mobile node moved to the new network. Therefore the IP signaling contributes the least to the overall delay. According to [18] to obtain the one-way transmission delay $D^s_{X,Y}$ of a message of size s in bits between node X and node Y the following equation can be used:

$$D^s_{X,Y} = \frac{1-q}{1+q} \cdot \left(\frac{s}{B_{\text{wl}}} + L_{\text{wl}} \right) + (d_{X,Y} - 1) \cdot \left(\frac{s}{B_{\text{w}}} + L_{\text{w}} + \varpi_q \right) \tag{9.24}$$

where q is the probability of wireless link failure, φ_q is the average queueing delay at each router between X and Y, B_{wl}, B_{w} is the bandwidth and L_{wl}, L_{w} is the delay of wireless and wired link, respectively. The values used for this analysis are given in Table 9.4.

Consequently, the handover delay for MIPv6 using the optimistic DAD concept and without performing the return routability process is:

$$D^{\text{MIPv6}} = D_{\text{L2}} + D^{\text{RtSol}}_{\text{MN,AR}} + D^{\text{RtAdv}}_{\text{AR,MN}} + D^{\text{BU}}_{\text{MN,HA}} + D^{\text{encap}}_{\text{HA,MN}}. \tag{9.25}$$

If the route optimization procedure is used and assuming that ($d_{\text{MN,HA}}$ + $d_{\text{HA,CN}}$) is greater than $d_{\text{MN,CN}}$, then the handover delay is:

$$\begin{aligned} D^{\text{MIPv6}_{\text{R}}\text{O}} = D^{\text{MIPv6}} + D^{\text{HoTI}}_{\text{MN,HA}} + D^{\text{HoTI}}_{\text{HA,CN}} \\ + D^{\text{HoT}}_{\text{CN,HA}} + D^{\text{HoT}}_{\text{HA,MN}} + D^{\text{BU}}_{\text{MN,CN}}. \end{aligned} \tag{9.26}$$

Taking into account HMIPv6, the handover delay is:

$$D^{\text{HMIPv6}} = D_{\text{L2}} + D^{\text{RtSol}}_{\text{MN,AR}} + D^{\text{RtAdv}}_{\text{AR,MN}} + D^{\text{BU}}_{\text{MN,MAP}} + D^{\text{encap}}_{\text{MAP,MN}}. \tag{9.27}$$

In the case of predictive FMIPv6 the handover delay is reduced to:

$$D^{\text{FMIPv6}} = D_{\text{L2}} + D^{\text{UNA}}_{\text{MN,nAR}} + D^{\text{encap}}_{\text{nAR,MN}}. \tag{9.28}$$

In the case of PMIPv6 the handover delay is:

$$D^{\text{PMIPv6}} = D_{\text{L2}} + D^{\text{PBU}}_{\text{MAG,LMA}} + D^{\text{encap}}_{\text{LMA,MAG}} + D^{\text{data}}_{\text{MAG,MN}}. \tag{9.29}$$

Finally, regarding the PFMIPv6 protocol:

$$D^{\text{PFMIPv6}} = D_{\text{L2}} + D^{\text{encap}}_{\text{nAR,MN}}. \tag{9.30}$$

After obtaining the handover delay for each protocol, the disruption time imposed by the handover process during the interval between consecutive sessions can be extracted utilizing the fluid flow mobility model. Similar to Sect. 9.4.1 the disruption time for a global mobility protocol "GP" is:

$$\text{DT}^{\text{GP}} = E(N_c) \cdot D^{\text{GP}}, \tag{9.31}$$

whereas for a local mobility protocol "LP" is given by:

$$\text{DT}^{\text{LP}} = E(N_s) \cdot D^{\text{LP}} + E(N_d) \cdot D^{\text{MIPv6}}. \tag{9.32}$$

Figure 9.15 shows the disruption time imposed by handover latency between sessions as a function of SMR. It can be seen that as the SMR increases (i.e. mobility

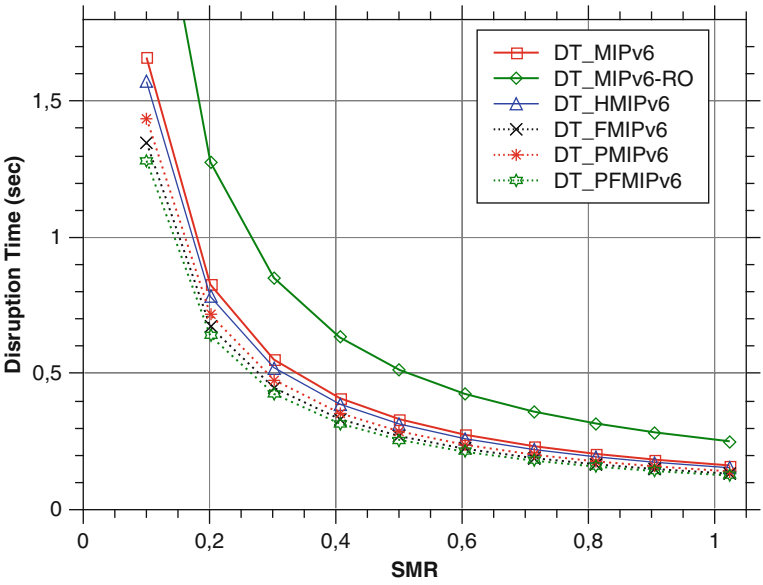


Fig. 9.15 Impact of SMR on disruption time

rate decreases) the disruption time declines too. Route optimization in MIPv6 has the highest values of disruption time due to the time needed to complete the return routability process. Conversely, the fast handovers schemes expedite significantly the handover delay and the consequent disruption time. This is the result of the fact that the data traffic is redirected to the mobile node's new location before it moves there, delivering packets to the mobile node as soon as its attachment is detected by the new access router. Once again, network-based solution (i.e. PFMIPv6) has the best performance as it does not introduce any signaling between the mobile node and the access network. Finally, the pure local mobility protocols' performance is in between the one of the basic and the fast handovers. As explained in Sect. 9.2.2 the binding update in the case of intra-domain handovers is executed locally minimizing the overall delay.

9.5 Simulation Results

In order to evaluate the performance of IP mobility for live 3D video streaming, MIPv6 and PMIPv6 will be compared through simulation. For this purpose, two new modules were integrated in the basic version of Network Simulator 2 (ns2) [28]. The first [9] is an open source tool that enables the simulation of scalable video coding (SVC) transmission and the second [32] implements the

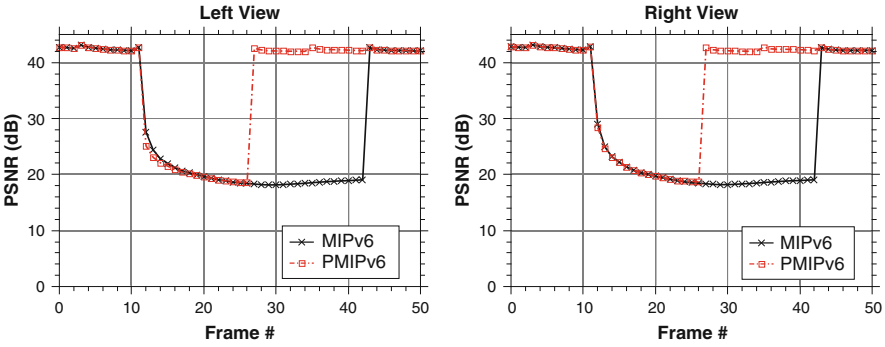


Fig. 9.16 PSNR comparison of MIPv6 and PMIPv6

Proxy Mobile IP protocol in ns2. The IEEE 802.11g WLAN was used as the radio access technology with a data rate of 54 Mbps, while the network topology was the one described in Sect. 9.4. The simulation results were obtained using the test sequence "Martial Arts", which is a high motion sequence with texture variation and standing camera, at full High Definition resolution of 1, 920 $\times$1, 080 pixels and 25 frames per second. The sequence was encoded using Medium Grain Scale scalability producing two layers with QP values (30,36) and an Intra-frame period of 8 frames. Finally, the playout buffer was set to a constant value of 500 ms.

Figure 9.16 shows that in the case of MIPv6 handover the drop of PSNR lasts for 30 frames while in PMIPv6 lasts only for 15 frames. The improved performance of PMIPv6 is a result of its fastest handover procedure that achieved by performing the registration procedure locally and without involving the mobile node. Additionally, from the simulations results it is obtainable that in the case of MIPv6, the handover delay is 1,172 ms while it is only 444 ms for PMIPv6.

It is worth noting that although the handover process is three times faster in the case of PMIPv6 compared to MIPv6, the PSNR drop lasts only half the number of frames. This is because, in the case of PMIPv6, the first frames received after the handover are B-frames (i.e. Frames #24 to #26) which cannot be decoded as the previous I-frame has been lost. On the other hand, in the case of MIPv6 the first frame received after the handover is an I-frame and it is decodable. As a result, it can be inferred that the disruption time may last more than the handover delay. In particular, assuming a rate of 25 frames per second, a GOP size of 8 frames and uniformly distributed handover occurrences inside a GOP, the disruption time will last 160 ms more than the handover delay on average. Additionally, the 26th frame which was transmitted distorted immediately after the completion of the handover process is shown in Fig. 9.17 for PMIPv6. For comparison, the first I-frame after the completion of the handover process is shown in Fig. 9.18a for MIPv6 and in Fig. 9.18b for PMIPv6. Eventually, the simulations results showed that the disruption time can be higher than the handover delay due to video coding characteristics.

Fig. 9.17 Distorted frame #26 for PMIPv6

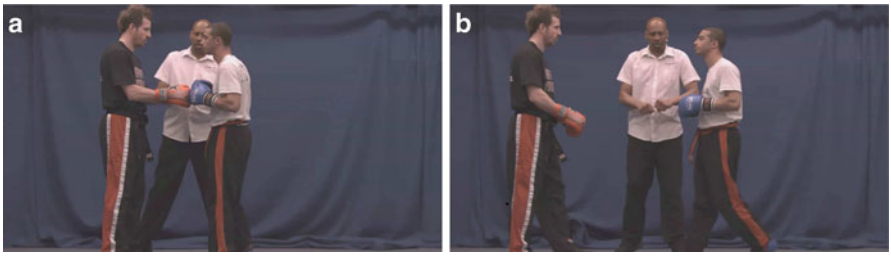


Fig. 9.18 (**a**) First I-frame for MIPv6, (**b**) first I-frame for PMIPv6

9.6 Conclusions

This chapter was focused on IP mobility management to support 3D video streaming services. Initially, the MIP-based protocols were briefly described as well as the IEEE 802.21 MIH protocol. Furthermore, an analytical comparison was given in terms of signaling cost, packet delivery cost, and disruption time. As analytical results showed in Sect. 9.4, local mobility reduces both the signaling cost and the handover delay but it cannot support route optimization. On the other hand, fast handovers technique minimizes the handover delay and can adopt route optimization in exchange for signaling cost.

Additionally, the simulation provided results showed that the recently introduced PMIPv6 outperforms MIPv6 for 3D video streaming. Through simulation results it was also shown that although the handover process is completed the video quality is not restored until the reception of a new I-frame. As a result, IP mobility management must focus on eliminating packet loss and packet end-to-end delay in order to fulfill the goal of seamless service provision.

Finally, in this chapter a new protocol for MIH based on subspace coding was proposed. The main aim is to make the original MIH protocol reliable in order to

reduce losses, retransmission load, and energy consumption. The protocol described by IEEE 802.21 has been adapted to include coding operations: a new stack and a new acknowledgement mechanism have been designed. The reference model of the protocol and the frame structure have been modified to make the proposed changes easy to deploy in IEEE 802.21 implementations.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements 264759 (GREENET) and 287896 (ROMEO).

References

1. Ahlswede R, Cai N, Li SYR, Yeung RW (2000) Network information flow. IEEE Trans Inf Theory 46(4):1204–1216. doi:10.1109/18.850663
2. Baumann FV, Niemegeers IG (1994) An evaluation of location management procedures. In: International conference on universal personal communications, San Diego, CA. doi:10.1109/ICUPC.1994.383136
3. Bernardos CJ, Kozat UC, Widmer J, Zorzi M (2012) Video over mobile networks [guest editorial]. IEEE Trans Vehicular Technol 61(2):871–876. doi:10.1109/TVT.2011.2178277
4. Bernardos C, Kozat U, Widmer J, Zorzi M (2013) Video over mobile networks [guest editorial]. IEEE Netw 27(2):6–7. doi:10.1109/MNET.2013.6485089
5. Cai N, Yeung RW (2006) Network error correction, part II: lower bounds. Commun Inform Syst 6(1), 37–54
6. Chou P, Wu Y, Jain K (2003) Practical network coding. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.697
7. Ho T, Médard M, Koetter R, Karger D, Effros M, Shi J, Leong B (2006) A random linear network coding approach to multicast. IEEE Trans Inf Theory 52(10):4413–4430 doi:10.1109/TIT.2006.881746
8. Institute of Electrical and Electronics Engineers (2009) "IEEE Standard for Local and metropolitan area networks—Part 21: Media Independent Handover Services". LAN/MAN Standards Committee of the IEEE Computer Society, pp c1-301, 21 Jan 2009
9. Ke CH (2012) myEvalSVC: an integrated simulation framework for evaluation of H.264/SVC transmission. KSII transactions on Internet and information systems, January 2012
10. Khirallah C, Vukobratovic D, Thompson J (2012) Performance analysis and energy efficiency of random network coding in LTE-advanced. IEEE Trans Wirel Commun 11(12):4275–4285. doi:10.1109/TWC.2012.102612.111380
11. Koetter R, Kschischang FR (2008) Coding for errors and erasures in random network coding. IEEE Trans Inf Theory 54(8):3579–3591. doi:10.1109/TIT.2008.926449
12. Koetter R, Médard M (2003) An algebraic approach to network coding. IEEE/ACM Trans Netw 11(5):782–795. doi:10.1109/TNET.2003.818197
13. Koodli R (2009) Mobile IPv6 fast handovers. RFC5568. Internet engineering task force (IETF), July 2009
14. Koodli R, Perkins C (2007) Mobile IPv4 fast handovers. RFC4988. Internet engineering task force (IETF), October 2007
15. Larsson P, Smida B, Koike-Akino T, Tarokh V (2013) Analysis of network coded HARQ for multiple unicast flows. IEEE Trans Commun 61(2):722–732. doi:10.1109/TCOMM. 2012.121112.110202
16. Li SYR, Yeung RW, Cai N (2003) Linear network coding. IEEE Trans Inf Theory 49(2):371–381. doi:10.1109/TIT.2002.807285

17. Li Z, Luo Q, Featherstone W (2010) N-in-1 retransmission with network coding. IEEE Trans Wirel Commun 9(9):2689–2694. doi:10.1109/TWC.2010.070710.090136
18. Makaya C, Pierre S (2008) An analytical framework for performance evaluation of IPv6Based mobility management protocols. IEEE Trans Wirel Commun 7:972–983. doi:10.1109/TWC.2008.060725
19. McEliece RJ (2003) The Guruswami-Sudan decoding algorithm for Reed-Solomon codes. Technical report, IPN progress report
20. Narten T, Nordmark E et al (2007) Neighbor discovery for IP version 6 (IPv6). RFC4861. Internet engineering task force (IETF), September 2007
21. Perkins C (2010) IP mobility support for IPv4, revised. RFC5944. Internet engineering task force (IETF), November 2010
22. Perkins C, Johnson D (2001) Route optimization in mobile IPv4. DRAFT. Internet engineering task force (IETF), September 2001
23. Halperin D, Greenstein B, Sheth A, Wetherall D (2010) Demystifying 802.11n power consumption. In: Proceedings of the international conference on power aware computing and systems, HotPower'10, Vancouver, BC, Canada. USENIX Association, Berkeley, CA. http://dl.acm.org/citation.cfm?id=1924920.1924928
24. Silva D, Kschischang FR, Koetter R (2008) A rank-metric approach to error control in random network coding. IEEE Trans Inf Theory 54(9):3951–3967. doi:10.1109/TIT.2008.928291
25. Soliman H, Castelluccia C et al (2008) Hierarchical mobile IPv6 (HMIPv6) mobility management. RFC5380. Internet engineering task force (IETF), October 2008
26. Sundararajan JK, Shah D, Médard M, Mitzenmacher M, Barros J (2009) Network coding meets TCP. In: IEEE INFOCOM 2009, Rio de Janeiro, Brazil, pp 280–288. ISSN 0743-166X. doi:10.1109/INFCOM.2009.5061931
27. Sundararajan J, Shah D, Médard M, Jakubczak S, Mitzenmacher M, Barros J (2011) Network coding meets TCP: theory and implementation. Proc IEEE 99(3):490–512 doi:10.1109/JPROC.2010.2093850
28. The Network Simulator - ns-2. [Online] http://nsnam.isi.edu/nsnam/index.php
29. Tran T, Nguyen T, Bose B, Gopal V (2009) A hybrid network coding technique for single-hop wireless networks. IEEE J Sel Areas Commun 27(5):685–698. doi:10.1109/JSAC.2009.090610
30. Yeung RW, Cai N (2006) Network error correction, part I: basic concepts and upper bounds. Commun Inf Syst 6:19–36
31. Yokota H, Chowdhury K et al (2010) Fast handovers for proxy mobile IPv6. RFC5949. Internet engineering task force (IETF), September 2010
32. Zhu Z, Wakikawa R, Zhang L (2009) Proxy mobile Ipv6 in ns-2.29. [Online] http://commani.net/pmip6ns, February 2009
33. Zhu Z, Wakikawa R, Zhang L (2011) A Survey of mobility support in the Internet. RFC6301. Internet engineering task force (IETF), July 2011
34. Luby MG, Mitzenmacher M, Shokrollahi MA, Spielman DA (2001) Efficient erasure correcting codes. IEEE Trans Inf Theory 47(2):569–584. doi:10.1109/18.910575

Part III
QoE and QoS Advances for 3D Media

Chapter 10
Dynamic QoS Support for P2P Communications

Evariste Logota, Hugo Marques, Jonathan Rodriguez, Fernando Pascual Blanco, Manuel Nuñez Sanz, and Ignacio Digón Escudero

Abstract Scalable Quality of Service (QoS) control is of paramount importance to effectively enable a seamless convergence of the rapidly evolving Peer-to-Peer (P2P) overlay communications over the Internet since the latter only supports best-effort service paradigm. For example, the European Union (EU) funded ROMEO project is focusing on a joint use of DVB-T2 and P2P overlay networks for live multimedia content sharing and collaboration among multiple users. This raises a strong need that the media packets transmitted through the P2P overlay delivery system must arrive earlier enough at the end users to assure a proper synchronization of the multiple views that may be received via the hybrid network. For this purpose, the P2P network must assure a certain QoS guarantee in terms of bandwidth, delay, jitter, and loss. More importantly, the control must be scalable to prevent excessive signalling and the related processing overhead, usually suffered in the traditional per-flow QoS control approaches. In this view, recent research effort claimed that the Internet resources can be efficiently over-provisioned (booking more resources in-advance) in such a way to allow differentiation of QoS control with reduced signalling overhead and increased resource utilization. This approach, however, needs further investigations for proper integration into innovative networking architectural designs to achieve performance. In addition to that, and in order to provide an end-to-end QoS, a mechanism to enforce prioritization policies within the customer's access network is also needed.

E. Logota (✉) • J. Rodriguez
Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal
e-mail: logota@av.it.pt; jonathan@av.it.pt

H. Marques
Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal
e-mail: hugo.marques@av.it.pt

F.P. Blanco • M.N. Sanz • I.D. Escudero
Telefónica I+D, Madrid, Spain
e-mail: fpb@tid.es; mns@tid.es; idigon@tid.es

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_10, © Springer Science+Business Media New York 2014

Hence, this chapter proposes a cross-layer control architecture that takes advantage of the Internet resources over-provisioning and the QoS policy enforcement within access networks to facilitate rapid development of P2P applications. The design aims to alleviate the requirements of buffers and the need for adaptation on end users' devices, thus allowing for cost-effective and rapid development of attractive services in similar hybrid content delivery networks.

10.1 QoS Control Requirements and Challenges in P2P Networks

The Information Technologies (ITs) have become integral part of our society with many evolving applications (e.g., data, multimedia, haptics, and personalized services), serving all kinds of activities whether professional, leisure, safety-critical applications, or business. As the citizens start to realize the benefits that the IT can offer, they are willing to select from a wide range of options to get the content they want ubiquitously anytime and over any facilities available. This imposed urgent needs to ease the development of attractive and innovative services that the end users can enjoy such as real-time interactive applications. However, the traditional networking designs have shown serious limitations in dealing with these expectations, which motivated a consensus to integrate all services over packet-switched infrastructures, using the Internet Protocol (IP), leading to the Next Generation Network (NGN) paradigm [1]. This way, the NGN could inherit control flexibility features and economic benefits of the Internet including its broad deployment to support innovations.

Nonetheless, the service convergence over the packet-switched networks raises major challenges since the Internet does not provide any kind of support for service differentiation while each service has its own requirements (e.g., in terms of bandwidth, delay, jitter, and packet loss). For example, Voice over IP (VoIP) applications requires 150 ms of (mouth-to-ear) delay, 30 ms of jitter and no more than 1 % packet loss [2]. The interactive video or video conferencing streams embed voice call, and thus have the same service level requirements as VoIP. In contrast, the streaming video services, also known as video on-demand, have less stringent requirements than the VoIP due to buffering techniques usually built in the applications. Other services such as File Transfer Protocol (FTP) and e-mails are relatively noninteractive and drop-insensitive. Moreover, the networking control solutions such as IP routing and the network management protocols, do need appropriate bandwidth guarantees to assure that control messages are correctly delivered on time to prevent performance degradation.

Besides, the communication paths usually correlate by sharing links inside a network. To ease the understanding, Fig. 10.1 presents a network control domain (e.g., a Differentiated Service (DiffServ) Domain [3]) with 3 Ingress Routers (IRs), 3 Egress Routers (ERs), and 7 core routers with a certain number of communication paths or multicast trees such as Path1, Path2, and Path3 originated at the IR1, Path4

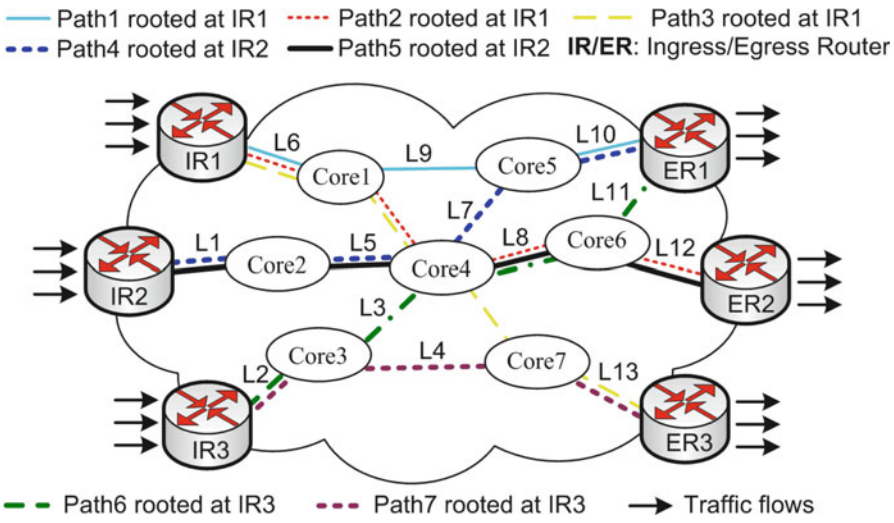


Fig. 10.1 Network scenario illustrating the requirements and challenges for QoS and resources control

and Path5 originated at IR2, Path6 and Path7 originated at IR3. In this scenario, the network links (e.g., L1, L2, L5, and L8) happen to be used by several paths originated at the same ingress router or different ingress routers. For example, the link L8 is used by three paths originated each at different ingress routers. Hence, one can imagine that, in real scenarios, the traffic flows passing through a given network interface must struggle to obtain the bandwidth resource that they need since the interface's capacity is shared and flows' behaviors are unpredictable. Therefore, the Internet resource must be properly controlled to effectively allow for the service convergence in the NGN due to the large amount of data involved in the communications, the heterogeneity of traffic characteristics, the users' terminal requirements as well as users' context such as preferences and location.

Another challenge is that, the rapidly evolving P2P applications promise to rely on the Internet infrastructures to provide various types of overlay services with both real-time and non-real-time requirements. For example, the European funded ROMEO project [4] is focusing on simultaneous delivery of live 3D media through both DVB-T2 and P2P transport technologies to facilitate application scenarios such as social TV, entertainment, and high quality real-time collaboration. To this, ROMEO aims to capitalize on the emergence of multiple views and scalable coding of rich content to transmit a base layer of view via DVB-T2 and multiple enhancement layers of views would be selectively delivered via P2P network. Given that the packets transmitted through the DVB-T2 would experience a relatively small delay, the packets transmitted through the P2P networks must also be received within acceptable delay for proper synchronization of the full view at end users in similar hybrid communication systems.

In this scope, the Internet Quality of Service (QoS) and resource control technologies consist of defining tools and techniques to provide predictable, measurable, and differentiated levels of quality guarantees to applications according to their characteristics and requirements by managing network bandwidth, packet delay, jitter, and loss parameters. To enable QoS support in the Internet, resource reservation and admission control have been investigated for many years as fundamental functions in networking control designs. These approaches involve control states maintenance and signalling messages to enforce the QoS policies (e.g., by configuring the schedulers [5, 6]) on nodes along the communication paths that media packets will follow.

The Internet Engineering Task Force (IETF) developed the Integrated Services (IntServ) [7], a QoS control architecture to provide end-to-end QoS support for each service individually over the Internet. In particular, IntServ guarantees QoS for each flow by explicitly reserving the requested amount of resource for the flow at every node on the path that the flow will take from its source to its destination, usually resorting to the Resource Reservation Protocol (RSVP) [8]. Upon receiving a request, the network is first signalled to probe the available resources. In case there is sufficient available resource, the network is signalled again so that the requested resource is reserved and the related states are maintained on all nodes on the relevant path. Hence, the reservation is released upon signalling when the session terminates. However, such per-flow states and signalling operations in IntServ have been severely criticized due to lack of scalability suffered under excessive control state and signalling overhead [9].

As alternative to IntServ, IETF introduced the DiffServ [3], a Class of Service (CoS)-based QoS architecture standard for the Internet. In class-based networks, traffic flows are classified into a set of CoSs at network borders (e.g., ingress routers) or at central stations (e.g., Bandwidth Brokers), according to predefined policies in terms of QoS, protocols, application types, etc. This way, the per-flow control states is maintained at the network edge while they are aggregated in the core nodes to reduce the overhead. For further scalability in the legacy DiffServ design, the network resource is assigned to each CoS in a static manner (e.g., a CoS is allocated a percentage of the link capacity) with no dynamic resource reallocation functions. Nonetheless, static resource allocation has clear limitations with poor resource utilization since traffic demands are dynamic and mostly unpredictable. Therefore, class-based network resources reservation must be carried out dynamically by taking network current resource conditions and the changing traffic requirements into account to improve resource utilization.

However, class-based control driven by per-flow signalling to increase the reservation of CoSs on per-flow demand basis as in [10], introduces unacceptable signalling and processing overhead which can easily overwhelm the Input/Output interfaces of nodes. Hence, the standard aggregate resource reservation [11] protocol introduced by IETF aims to allow for dynamically reserving more resources than a CoS needs, so that both QoS states and signalling overhead can be reduced. This approach has been researched for many years [12–14] as promising method to achieve differentiated QoS in a scalable manner. Prior and Sargento [15] provided valuable studies and analyses

of both per-flow and the standard aggregate reservation approaches. They found out that, per-flow approach allows high utilization of bandwidth at the price of undue signalling overhead and thus fails to scale. On the other hand, the standard aggregate resource control approach reduces the signalling overhead to scale at the expense of low resource utilization; the more over-reserved resources, the more the signalling overhead decreases, and more resources are wasted. It is also deeply investigated in [16] that aggregate over-reservation imposes a strong trade-off between the reduction of signalling overhead and the waste of resources.

While aggregate resource over-reservation has proved to achieve scalability with QoS guarantee, the waste of resources usually placed by inefficient control of residual bandwidth (over-reserved but unused) among CoSs has made network operators reluctant. In this respect, recent proposals [12] demonstrated that it is possible to avoid the waste of resources while using over-reservation in dynamic systems. Therefore, this chapter aims to provide a scalable *Internet Resource and Admission Control Subsystem* (IRACS) QoS control schemes which relies on efficient aggregate resource over-reservation techniques [17] in such a way to support P2P communications with significant reduction of QoS control signalling overhead while assuring differentiated control without wasting resources. This intends to enable the IP networks with flexible and cost-effective control mechanisms to support transparent service convergence, which is of paramount importance to motivate the success of added-value and innovative applications such as the ROMEO-alike services (e.g., real-time and bandwidth consuming) at reasonable cost.

In addition, the access network is a segment that sometimes represents a bottleneck when deploying services over the Internet. The main problem here is that when customers use several services competing through the same access network, such as media streaming, web browsing, file transfers, online playing, system updates, etc., they tend to exhaust access network resources, causing services to not be delivered properly. In some scenarios this problem can be fixed by provisioning static prioritization policies within the access network elements, but this approach fails when dealing with P2P services like ROMEO because of its dynamic nature. Therefore, a solution to dynamically enforce QoS policies within the access network will be presented and will be based on the 3GPP Policy, Charging and Control (PCC) [18] architecture. That architecture will let P2P services to request the desired flow prioritizations to the access network in a simplified way and just in the moment when the service is going to be delivered through the access network.

This chapter is organized as follow. The Sect. 10.1.2 describes a QoS over-provisioning architecture for P2P networking and the Sect. 10.1.3 details a priority-based QoS management architecture within access networks. Further, the Sect. 10.1.4 presents the ROMEO end-to-end QoS control approach and the Sect. 10.1.5 concludes the chapter.

10.2 A QoS Over-Provisioning Architecture for P2P Networking

The main objective of this section is to describe a scalable QoS architecture to enable bandwidth-aware IP transport for P2P communications over the Internet infrastructures, with low QoS reservation control signalling and the related overhead (e.g., processing load and energy requirements). To facilitate the understanding, Fig. 10.2 is used to illustrate a heterogeneous networking scenario in which rich 2D and 3D contents, destined to interested client peers, are delivered simultaneously through a DVB-T2 delivery system and a P2P delivery network. In particular, the P2P transport system encompasses three access domains such as the *Access Network* (AN) A, the AN B, and the AN C. In each AN, the client peers are expecting to enjoy live attractive services on their multihomed devices connected to the networks. These ANs are attached to a core network infrastructure under a central control entity called *Network Control Decision Point* (NCDP). The NCDP is placed in a service/control network domain together with a set of dedicated content processing and delivery servers. As shown in Fig. 10.2, these servers include a media encoding server to process content so as to provide scalable codecs to the ROMEO server which is responsible for *Authentication*, *Authorization, and Accounting* (AAA) functions. Besides, a super-peer server is deployed inside the service/control domain as the responsible for building several P2P overlay multicast trees and managing the overlay topology for the P2P content delivery. As it is detailed in Sect. 10.1.4, a super-peer is deployed per *Internet Service Provider* (ISP) for the sake of scalability of the P2P overlay system. In this chapter, a border node that connects an access or service/control network to the core is called *Edge Router* (ER) while the core represents a backbone infrastructure to bridge connectivity between different network domains.

In this scenario (see Fig. 10.2), the ROMEO server transmits a 3D view (a base stereoscopic view) through the DVB-T2 system and other 3D views such as enhancement views are delivered over the P2P network as relayed by the super-peer server. Thus, the DVB-T2 technology is used to provide a seamless content with a base view streaming to all end users within the broadcast area. Then, the views transmitted over the P2P system are expected to take into account several control metrics such as end users' preferences, locations, and devices capabilities. Therefore, the hybrid system envisages media distributions among collaborating peers with support for QoS as well as service personalization. Hence, it is mandatory that the streams received from DVB-T2 and the P2P network must be synchronized to provide acceptable quality of the full multi-view 3D content to individuals as well as to the collaborating peers. However, the P2P communications depend on the underlying Internet infrastructure which only supports best-effort service paradigm as we referred earlier. Therefore, a proper QoS support must be carefully engineered to take the P2P overlay resource capabilities and the underlay network resource availability into account simultaneously to allow for service convergence over the Internet while improving the network overall utilization, the so-called cross-layer control approach which is depicted in Fig. 10.3.

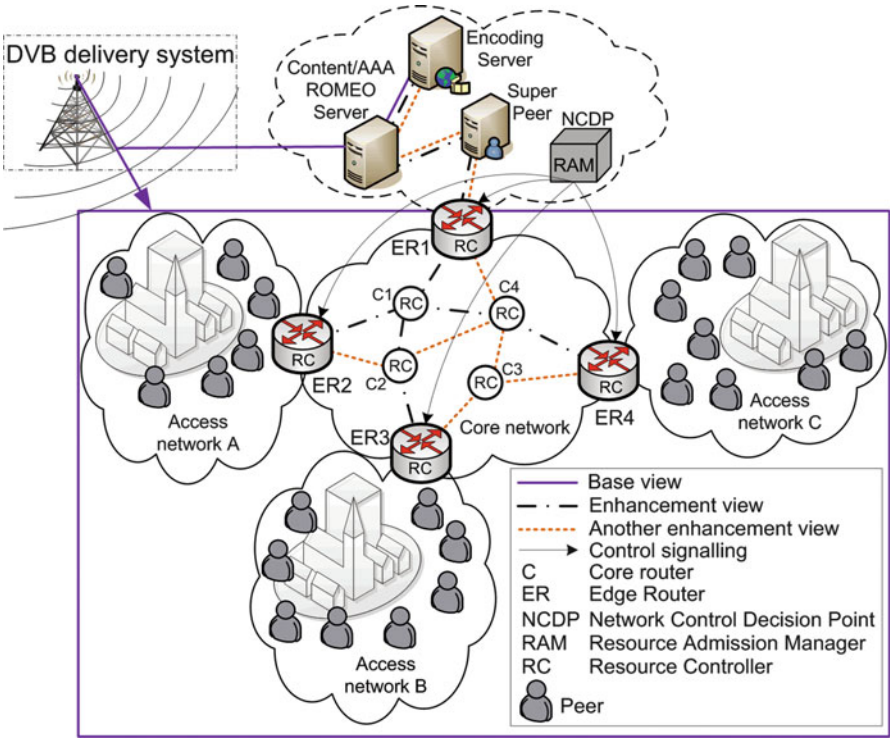


Fig. 10.2 A use case architecture for ROMEO networking scenario

In Fig. 10.3, the NCDP embeds a *Resource and Admission Manager* (RAM) as being the responsible for defining proper control policies and managing the distribution or access to the underlay network resources. Besides, the network nodes (e.g., routers) implement a Resource Controller (RC) that enables them to enforce the control decisions based on the instructions that they receive from the RAM. The physical mapping of RAM and RC to server and to network nodes respectively is shown in Fig. 10.2 and the related functional architecture is provided in Fig. 10.3. It is also worth mentioning that, the overall mechanism deployed to coordinate the control of the RAM and the RC is referred to as the IRACS. In particular, the RAM includes several functional modules such as a Resource Reservation Subsystem (RRS), a Control Information Base (CIB), an Admission and Control Subsystem (ACS), and one interface for communications with external nodes. Besides, each network node implements a Resource Controller (RC) with added-value functions to allow for the enforcement of the control decisions taken by the RAM.

Moreover, Fig. 10.3 illustrates the P2P overlay subsystem which is hosted by the super-peer shown in Fig. 10.2. Further details on these modules in terms of their relevant functions and interactions are provided in the following subsections.

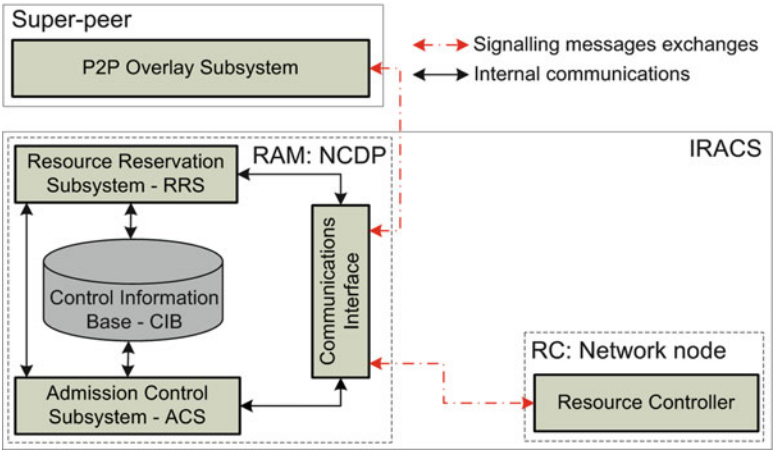


Fig. 10.3 Cross-layer QoS control architecture

10.2.1 Resource and Admission Management Functions: RAM

The RAM implemented in the IRACS approach is responsible for network access and resource allocation to support IP transport with heterogeneous QoS requirements. In other words, the RAM controls the operators' network infrastructures by granting or denying access to the related resource consumption in a way to improve the network utilization while guaranteeing differentiated QoS for all admitted sessions based on CoSs. Such differentiation of QoS control is a must for seamless convergence of all types of services over the Internet. Hence, the RAM is responsible for defining appropriate control policies and dictating them for the enforcement on network nodes to prevent some Internet applications (e.g., bandwidth demanding P2P communications) from starving other applications of resources.

The RAM also performs traffic load balancing to avoid unnecessary congestion occurrence in support for packet delay and jitter control. As such, it facilitates the synchronization operations of multiple views that end users receive from diverse technologies or multiple communication paths. Also, the RAM uses efficient resource over-reservation techniques to improve control scalability, that is, without overwhelming networks with QoS reservation signalling, states and processing overhead while guaranteeing Service Level Agreements (SLAs) in terms of bandwidth demands. The RAM achieves these functions by means of a good knowledge of the underlay network topology, the related link resource statistics and control through interactions between various modules such as the RRS, the CIB and the ACS which are respectively described in Sect. 10.1.2.1, 10.1.2.2, and 10.1.2.3.

In purely centralized networks, a single RAM will take overall control of the network as in Fig. 10.2. In hierarchical scenarios spanning multiple domains as described in Sect. 10.1.4, a RAM is deployed per core network for scalability reasons, and inter-domain connections are performed according to predefined

SLAs between the domains. In other words, IRACS aims to optimize the network utilization across the core domains by enforcing scalable QoS measures. In the access networks, the QoS is assured by means of the virtualization concept described in Sect. 10.1.3. This way, ROMEO is expected to guarantee end-to-end QoS provisioning over the P2P network without unnecessarily depriving the Internet background traffic of resources.

10.2.1.1 Resource Reservation Subsystem

In the RAM architecture, the RRS is exploited to create and manage bandwidth-aware multicast trees across the core network domain in a way to allow for connection between access networks and peers. The RRS module deploys aggregate resource over-reservation control scheme [12] to dynamically define appropriate policies for resource sharing among various CoSs on network interfaces upon need. Aggregate over-reservation means that a CoS may be reserved more bandwidth than it currently requires, according to the local control policies. This approach is used to prevent excessive QoS control signalling, states and processing overhead to achieve scalability and to reduce session setup time as well. Moreover, it enables RRS to avoid CoS starvation by means of proper readjustment of reservation parameters dynamically upon need, such that the performance can be achieved without increasing session blocking probability unnecessarily.

10.2.1.2 Control Information Base

The RAM uses the CIB to maintain a good knowledge of the underlying network topology and the related resources statuses. It stores the multicast trees created inside the network under the control of the RAM and the IDs of the outgoing interfaces that belong to the trees. Moreover, it maintains the overall capacity of each interface, the amount of bandwidth reserved and used in each CoS on the interface [17]. The CoSs configured on the interface are also maintained along with relevant information about the active sessions inside the network. An active session's information includes, but is not limited to, the bandwidth required by the session, the session ID, the IDs of the flows that may compose the session, the ID of the CoS to which the session belongs, the source ID, the destination ID, the ports IDs, associated multicast tree's ID, and the multicast channel.

10.2.1.3 Admission Control Subsystem

The ACS enables RAM to accept or reject service requests to a network, depending on the service requirements in terms of QoS (e.g., bandwidth) and the network resource availability reported by the CIB local database. Therefore, the RAM provides an interface for interactions with the P2P overlay control subsystem

embedded in the super-peer. It is worth mentioning that this interface allows also for receiving service demands from the ROMEO server or a peer upon need, depending on the overlay specific control mechanism and requirements. Whenever a session is admitted, terminated or the QoS requirements of a running session are readjusted in a CoS on a communication tree, the ACS process updates the resource utilization status such as the bandwidth usage in the concerned CoS on the related outgoing interfaces in the local CIB. Considering that resource is over-reserved throughout the network, the ACS is able to admit, terminate or readjust the QoS demands of several sessions without signalling the nodes inside the network as long as the over-reservation is not exhausted on the distribution trees involved in the process, leading to scalable admission and QoS control. When the over-reservation is exhausted on a tree, the ACS triggers the RRS to define new reservation parameters such as the amount of resource to be reserved and the reservation thresholds for the relevant CoSs along the tree as detailed in [12]. After the new reservation parameters have been successfully computed, the RAM conveys them to the nodes on the tree so that the new control policies are enforced on the nodes using the RC modules. This way, the reservation parameters are defined and readjusted dynamically in a way to prevent CoS starvation or unnecessary waste of resources while the QoS control signalling frequency is reduced for scalability. The RC functions are described in the subsequent subsection.

10.2.2 Resource Controller Functions: RC

The RC module implements basic control functions required in all network nodes and mainly operates in the routers at the ISP network as illustrated in Fig. 10.2. In particular, it deploys elementary transport functions to enable UDP port recognition (routers are permanently listening on a specific UDP port) or IP Router Alert Option (RAO) [19] on nodes to properly intercept, interpret, and process control messages. It interacts with Resource Management Functions (RMF) [20] to properly configure schedulers on nodes [5, 6], thus ensuring that each CoS receives the amount of bandwidth allocated to it to provide QoS-aware data transport across the network. For flexibility, it interfaces with legacy protocols (e.g., routing protocols, existing system databases) on nodes in order to improve performance. For example, the RC is able to exploit legacy control databases such as, but is not limited to, the *Management Information Base* (MIB), *Routing Information Base* (RIB) or *multicast Routing Information Base* (MRIB) and *Forwarding Information Base* (FIB), according to the control instructions received from the RAM.

When deployed at network border, the RC is enabled to learn inter-domain routing information from the traditional Border Gateway Protocol (BGP) [21] for proper packet delivery between various domains. It also interacts with traffic control and conditioning for traffic shaping and policing according to operator's local control policies to force admitted traffic flows to comply with the SLAs between network users and the providers, which functions are available in most of the Differentiated Services (DiffServ [3])-based frameworks. Further, the RC is used

to enforce multicast trees decisions upon receiving instructions from the RAM. It is also used to allow control messages to collect the IDs of outgoing interfaces and their capacities on trees as being *Record Route Object* (RRO) [22, 23]. When instructed by the RAM through a control message, the RC enables nodes to record the ID of the previous outgoing interface visited by the message so that asymmetric route issues can be avoided in reverse direction of trees [24].

As a result, the IRACS approach pushes network control complexity to the RAM at the NCDP server, and the core nodes are left simpler by implementing the RC for scalability reasons. In most of the cases, the RC implementation will not be needed at the peers, since the operator is providing resource reservation control service and peers are connected through the network. The RC module, however, can be implemented at peer level to enforce ROMEO application to guarantee QoS requirements when peers share the same broadcast domain.

10.2.3 Network Initial Configurations and Operations

At network initialization, when nodes boot up, the RAM, especially the ACS module, gets network topology information by importing such information from existing (resident) link state routing protocols [25]. The ACS then uses an appropriate algorithm (e.g., Dijkstra [25]) to compute all possible edge-to-edge routes inside the core network under its control. As in [13], a combination of the edge-to-edge routes leads to all possible edge-to-edges branched routes. Among these computed routes, the ACS selects the best routes that can be used for service delivery. A route can be selected based on but not limited to its number of hops and bottleneck outgoing interface capacity. It is worth mentioning that a bottleneck outgoing interface on a route is the outgoing interface which has the smallest capacity on the route. Then, the ACS allocates a unique multicast channel (S, G) for each selected route where S is the IP address of the edge router at which the route originates and G is the IP multicast address assigned to the concerned tree. Besides, the ACS triggers the RRS (through interface 1) and the latter defines initial over-reservation parameters to be enforced on interfaces inside the network. After that, the ACS encapsulates this information together with the route record object RRO in a control message and sends the message to the nodes on each route.

As the control message is travelling along a route, every visited router hosting the RC module intercepts the message and configures its local multicast routing table as well as the initial over-reservation parameters destined to its interfaces accordingly. Also, the control message is forced to follow the desired route by means of the route record RRO which enables source routing such that multiple QoS-aware multicast trees are thus initialized for use inside the network. In a pure centralized scenario, a single RAM maintains knowledge on the entire network and related trees. In a hierarchical control scenario, an end-to-end route may be a concatenation of trees from each of the domains that lie on the route as in Sect. 10.1.4 to ensure that packets are pinned to the desired routes where they receive the bandwidth reserved for them. Hence, every RAM in a domain properly maps traffic flows to its local trees

according to its local QoS control model, independently of the other RAMs and end-to-end control is assured in a scalable manner across heterogeneous network environment as further detailed in Sect. 10.1.4.

10.2.4 Cross-Layer QoS Control Using Resource Over-Provisioning

This subsection describes a cross-layer mechanism in which incoming overlay peers' QoS demands and the underlying network resource availability are efficiently taken into account simultaneously to allow for improving the overall network utilization and the perceived quality of the end users. In other words, as the network is initialized (see Sect. 10.1.2.3) and set to run, every authorized QoS-sensitive P2P session request must be jointly processed by the P2P overlay subsystem and the IRACS resource and ACS. Hence, each session request specifies its desired QoS (e.g., bandwidth and buffer) and the related traffic characteristics (e.g., flows IDs, source and destination IP addresses and ports). Then, the P2P overlay subsystem encapsulates this information in a *Next Step In Signalling* (NSIS) compliant protocol [17] or any other protocol specified by the operator, and sends it to the RAM through the interface 7 in Fig. 10.3. A session request process may be triggered by the ROMEO server, a super-peer, a peer, or another RAM in multiple RAMs control environment, depending on the scenario, as illustrated in Sect. 10.1.4.

To ease the understanding of the interactions between various network elements such as a joining peer, a super-peer, a RAM, and the ROMEO server for a session setup, Fig. 10.4 depicts a scenario in which a peer wants to enjoy a live 3D content service in the network topology presented in Fig. 10.2. Hence, the peer first interacts with the ROMEO server for Authentication, Authorization, and Accounting purposes. During this control phase, the peer and the ROMEO server exchange also the QoS requirements (e.g., bandwidth) of the desired media stream together with the related traffic characteristics (e.g., codec and peak rate). In case these AAA operations are successful, the server redirects the peer to the super-peer by providing the address of the latter. This is important to prevent unknown peers from affecting the media delivery performance.

Based on the instructions received from the server, the joining peer issues a connection request to the super-peer. The request carries the desired session's QoS and the related traffic characteristics. Hence, upon receiving the request, the super-peer contacts the NCDP server (which hosts the RAM) to request for the underlay network resources in order to establish QoS-aware connectivity for the media streaming. The request to the NCDP includes the IP address of the joining peer and that of the ROMEO server as well as the QoS requirements and the traffic characteristics. Thus, the NCDP selects appropriate multicast trees (preestablished as in Sect. 10.1.2.3) that can guarantee the requested QoS to connect the super-peer to the ROMEO server and to the joining peer. In case the trees' selection succeeds, the NCDP instructs the edge routers that lie on the selected trees to bind the

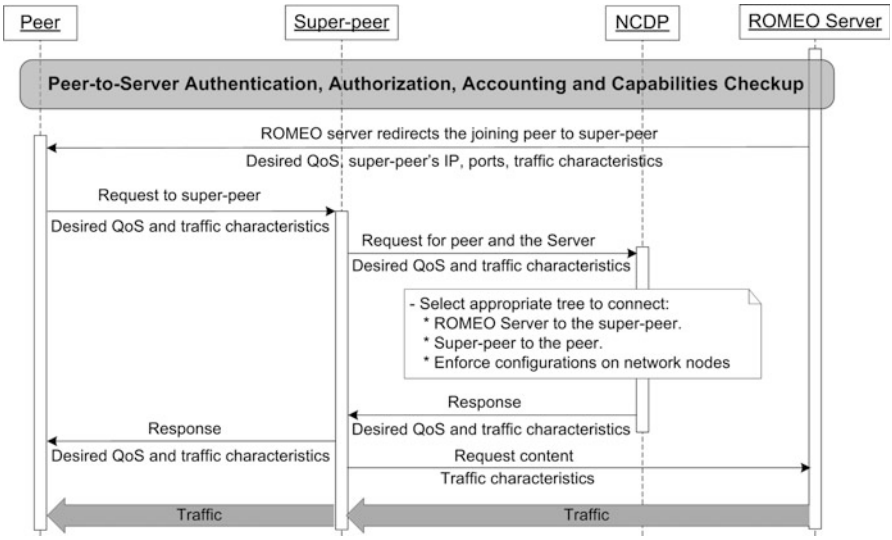


Fig. 10.4 A use case for P2P session setup sequence chart

incoming session parameters (e.g., QoS requirements, source' and destination' IP addresses, ports and transport protocol) with the selected trees. Thus, incoming media packets are correctly encapsulated at ingress edge routers to follow the desired trees, so they enjoy the QoS reserved for them, and the packets are decapsulated at relevant egress edge routers for delivery to the end user.

Considering that each network interface is initialized with extra resource reservation, several connection requests may be established without QoS reservation signalling as long as there is enough over-reservation available on the selected trees. As demonstrated in [12], the reservations of CoSs on a tree are readjusted only when the reservation of a CoS exhausts and the CoS is demanding more resources, depending on the incoming session requests.

10.3 An Architecture for Priority-Based QoS Management

A priority-based QoS solution is employed within a virtualization platform hosted at IP Edge node (e.g., see ER in Fig. 10.2) to ensure that the user perception meets the expected QoE for all services subscribed to the residential environment, while IRACS controls the QoS in the core networks (see Sect. 10.1.2). As depicted in Fig. 10.5, that solution will be provided by using two main functions: a Policy and Charging Enforcement Function (PCEF) and a Policy and Charging Rules Function (PCRF) [18]:

- The PCEF is the responsible for the QoS enforcement within the IP edge node. In that case the PCEF module will be embedded within the Broadband Remote Access

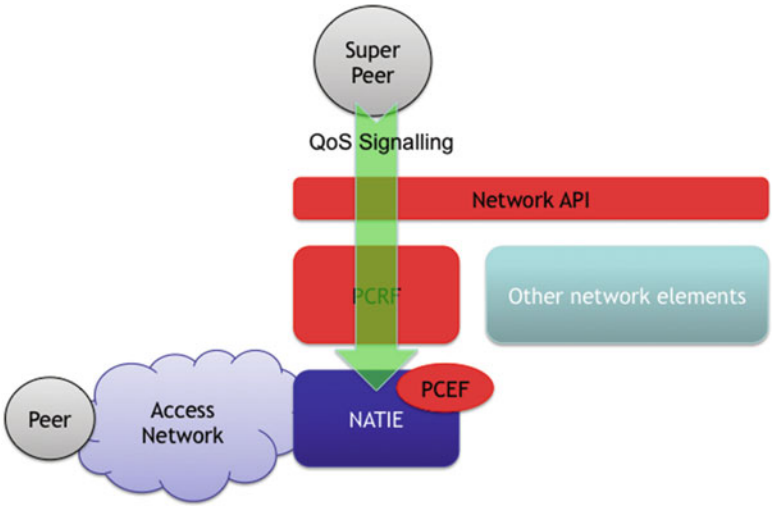


Fig. 10.5 Architectural design

Server (BRAS). In the virtualization use case, the BRAS becomes the Network Address Translation Implementer Equipment (NATIE) because it acquires functionalities typically placed within the Customer Premises Equipment (CPE).

- The PCRF is the responsible for defining the QoS rules that will be applied to a particular user. It receives from the upper service layers the physical parameters to identify a certain flow to be prioritized and communicates with the PCEF in order to enforce the rules in the access network.

Figure 10.5 depicts the designed architecture and places the described functions.

Being the super-peer node the element aware of the flows to be prioritized (because it is the one that is serving the content to the peer) it is the one selected as the element to notify the QoS needed towards the access network: to the PCRF through the Network API—an API that intends to simplify the interfaces towards the network to upper service layers. This notification will be passed to the policy manager (PCRF) through the Network API in order to enforce the prioritization needed to achieve the requested QoS within the NATIE.

10.3.1 The Network API

The Network API layer is deployed in order to ease the management of the access network resources to the upper service layers. It enhances the service provisioning, adding flexibility and real-time control over the flow prioritization, among other things. A unique interface is responsible of managing these features as requested, offering a transparent way to access any of the offered functions.

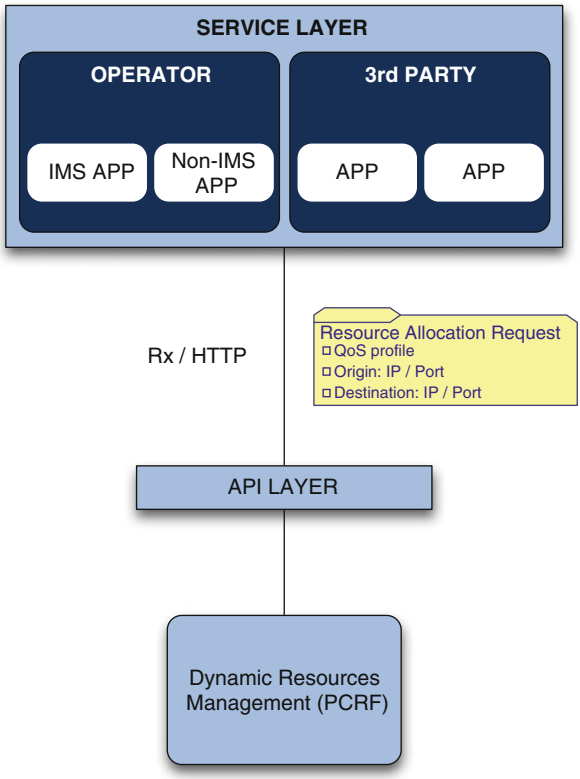


Fig. 10.6 Network API

The API layer offers a way to prioritize a specific flow within a given channel based on a pair of tuples (IP address/Port) describing the origin and destination of the desired flow. It is also necessary to choose the desired QoS profile, since there will be several profiles predefined for different services.

Figure 10.6 depicts how upper service layers owned by the service provider (IP Multimedia Subsystem (IMS) based or not) and third party applications (APPs) make use of the network capabilities by requesting resources to the PCRF through the Network API (API Layer).

10.3.2 Policy and Charging Function Control

The PCRF is the element in charge of the policy management within the access network infrastructure. It is in contact with the IP Edge node (in this case the NATIE) to apply specific policies to certain users, depending on the needs of every user. It provides a dynamic policy configuration and improves the classic static configuration of the access networks.

Figure 10.7 depicts how the QoS enforcement platform composed by the Network API, the PCRF and the PCEF connects with the Virtualization Platform. The

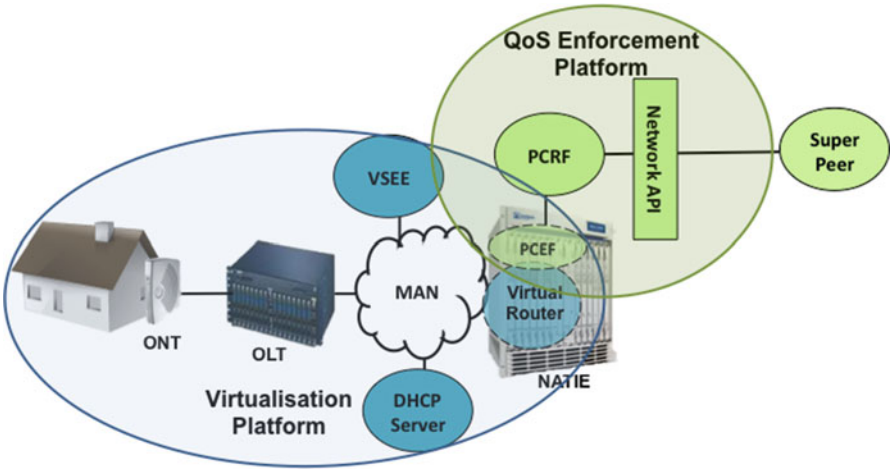


Fig. 10.7 PCRF

Virtualization Platform, which basically removes the classical layer 3 router from the customer premises and shifts its capabilities to the service provider premises, is composed by

- The Gigabyte Passive Optical Network layer 2 access network (placing an Optical Network Terminal within the customer premises and an Optical Line Terminal within the service provider premises).
- The Virtual Software Execution Environment (VSEE) in order to place virtualized services in a virtual machine for the customer.
- The Dynamic Host Configuration Protocol (DHCP) server in order to allocate IP addresses to the customer devices.
- The Metropolitan Area Network (MAN) in order aggregate several access network (GPON-based or not)
- The NATIE, the BRAS acquiring the layer 3 capabilities shifted from the customer premises to the service provider premises.

It is also shown that the PCRF enforces the desired policies to the PCEF, embedded within the NATIE.

The PCRF is composed by three elements. The first one is the Subscriber Profile Repository (SPR), which is the internal database where the users must be provisioned to be managed by the PCRF. The second one is the Multimedia Policy Engine (MPE), which is the policy manager module, and it is configured through the third one, the Manager (MGR), which is the configuration web interface.

10.3.2.1 Subscriber Profile Repository

The SPR contains the database where the users are provisioned. If the PCRF receives a message indicating a user started a session not provisioned within the SPR, there will not be a session created within the PCRF for that user, the user will access the network but no policy changes will be applied to that user.

Users are identified by using their Network Address Identifier (NAI), the E.164 [26] with the Mobile Station Integrated Service Digital Network (MSISDN) or the International Mobile Subscriber Identity (IMSI) [27], so at least one of these fields must be provisioned. In addition, every provisioned user must be provided with four fields:

- DefaultBW (Bandwidth by default)
- DefaultQoS (Quality of Service by default)
- InstalledBW (Bandwidth applied)
- InstalledQoS (Quality of Service applied)

These fields are verified and updated every time a policy is applied to a user. Bandwidth profiles define the upload and the download speed to be applied to the user (in kbps), and QoS profiles defines the flows to be prioritized. The fields that define a flow to be prioritized are:

- IP_DEST (destination IP address)
- PORT_DEST (destination port)
- IP_ORIG (source IP address)
- PORT_ORIG (source port)

10.3.2.2 Multimedia Policy Engine

The MPE is the policy manager engine. The entire configuration performed within the MPE must be done through the MGR, but it also exist a command line tool to configure it.

It is also possible to see all the internal sessions within the MPE. Every internal session represents a network user. When a user accesses to the network, an internal session is created (by using the information mentioned before in Sect. 10.1.3.2) within the MPE. That way by modifying user's internal session within the MPE there will be changes in their applied policies within the NATIE.

10.3.2.3 Manager

The MGR is a web tool to configure the MPE. It supports the configuration of several MPEs although in that case there will be only one.

Figure 10.8 shows how user's policy definition is performed within the "Policy Library" section, where policies can be created using blank ones, using a template or using an existing policy.

Fig. 10.8 MGR, policy definition

Fig. 10.9 MGR, logs

After that, the policy can be defined by using a number of conditions to be applied and actions to be executed. A normal action will be to install a particular service pre-configured within the NATIE by using a label identifying the service itself.

Figure 10.9 depicts how PCRF logs can be shown. It is possible to configure the search parameters in order to obtain only the desired results.

10.3.3 Architectural Design

This proposal has as a central key, the PCRF, controlling the configuration process by signalling (following the 3GPP PCC model [18] by using the Gx reference point) towards the NATIE. The following picture depicts the architecture:

However, if a network operator wants to deploy that architecture in the short term, due to current market limitations, there could be several drawbacks to achieve that goal architecture and any of them has an already available solution in the market that could help to achieve the objective. Following these lines and summing up:

- Currently it is easy to find a BRAS not supporting the Gx interface (defined by the 3GPP PCC model). In this case, the PCRF should support RADIUS CoA (RADIUS Change of Authorization), which is widely implemented in BRASs. Otherwise, the use of an AAA server is recommended if the PCRF does not supporn RADIUS CoA interface. The AAA server would act as Gx—RADIUS CoA translator.

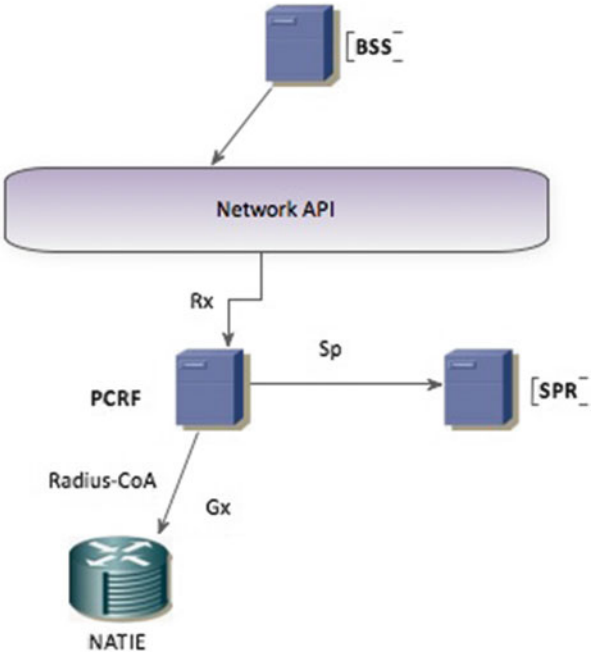


Fig. 10.10 Architectural design

- Finally, there is no a well defined network API running in the operating network. In this case, the service layer (in our case, the Super-Peer requesting the QoS to the access network) will have to be aware of the network complexity (its topology), and to implement complex database operations, as well as the Rx reference point instead of a simpler and widely used http-like interface as SOAP, or JSON [9].

In the ROMEO case, the NATIE will not support Gx interface, so it will be managed from the PCRF through RADIUS CoA (the selected PCRF obviously implements RADIUS CoA). A Network API avoids the upper service layers to be aware of the complicated network topology, and eases the development of applications and services able to manage dynamically network resources. The Network API will support a HTTP interface, so that service modules (such as the super-peer) can use it to enforce policies within the access network (NATIE) (Fig. 10.10).

10.3.4 Parameters Needed to Enforce the QoS

To guarantee a certain QoS to the ROMEO data flows, there are five parameters needed in order to identify each flow:

- Source IP address
- Destination IP address

- Source port
- Destination port
- QoS profile to enforce

These parameters will be provided to the PCRF through the network API from the appropriate ROMEO entity, at this time the most suitable candidate is the super-peer.

10.3.5 QoS Enforcement Within the Network Access Server

In ROMEO, the QoS in the access network relies on DiffServ protocol to ensure the proper treatment for different types of traffic. In this scenario, the NATIE is the entity in which QoS enforcement will be performed. Figure 10.11 shows the generic actions the NATIE triggers to incoming traffic.

Figure 10.9 blocks functionalities are defined as follows:

- *Filter*: The first action when a packet reaches the customer Virtual Router and Forwarding VRF (a virtual routing domain dedicated to a particular customer) within the NATIE is the identification of the type of service the packet corresponds to. Packets can be identified according to different factors (IP destination, IP origin, transport protocol used, etc.).
- *Traffic classifier*: This block assigns a forwarding class and Loss Priority parameter to the filtered traffic. According to the different time constraints services may be subjected to, the filtering action (see Fig. 10.11) divides traffic in several classes: Best Effort (BE), without QoS guarantees, Assured Forwarding (AF) that ensures four special different treatments without fixing any QoS guarantee (no SLA is established) and Expedited Forwarding (EF) that it is the class that offers better QoS guarantees (throughput, loss rate, delay, and jitter), being equivalent to a dedicated line.
- *Policier*: The application of Input/Output policies allows fixing the service bandwidth. In most cases, Input/Output policies are used to limit the bandwidth for the Internet service.
- *ReMarking*: Before the traffic is divided in queues, the DSCP [28] packet field is remarked to allow subsequent systems as the MAN, the GPON, or the services give the agreed treatment to each type.
- *Scheduler*: Finally the packets are placed in the scheduler. The scheduler has defined diverse queues according to the degree of prioritization envisaged to cover all services, being possible to configure different resource configuration to each queue such as bandwidth and buffer length.

Figure 10.12 shows a possible configuration of the scheduler. Four queues have been configured, where the lower the queue number is the more priority the traffic has. Queue number 1 is dedicated to expedited forwarding traffic and corresponds to maxim priority. This type of traffic is allocated to the priority queue so minimum bandwidth, maximum delay, and maximum jitter is guaranteed.

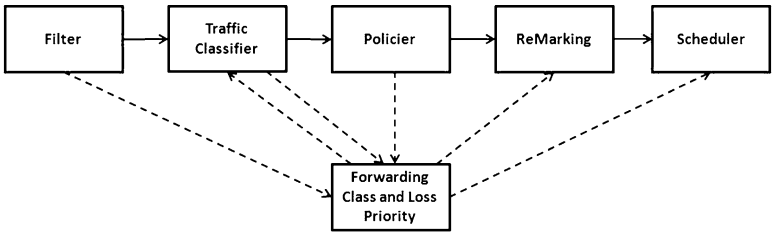


Fig. 10.11 QoS diagram block as performed by NATIE

Queues 2–4 are configured to operate with a weighted round robin algorithm. Queue 2 and queue 3 are configured to process the assured forwarding traffic. Finally queue 4 is devoted to process best-effort traffic. The weights allocated to each queue depend on the minimum bandwidth each type of service needs in order to meet its QoE. Figure 10.12 shows the configuration details applied to the example: (Table 10.1).

10.4 QoS-Enabled End-to-End Communications in P2P Networks

The main objective of this section is to describe how the ROMEO deals with QoS across multiple networks domains. To facilitate the understanding, Fig. 10.13 is used to illustrate the IRACS resource and admission control approach, involving several network domains. In particular, we assume that ISP Y is hosting the ROMEO server and a given peer A, in an ISP A's access network, are connected through a transit ISP T. Hence, we assume that the ROMEO AAA Server, and the Encoding Server are hosted in an ISP A's infrastructure which is connected to the ISP B's network via the Border Routers—BRs—(BR1 in ISP A and BR2 in ISP B). Notice that an Edge Router (ER) is a node that bridges connection between an access and a core networks while a BR is used to connect two core networks. Besides, the ISP B's network is composed of 2 access networks (Access B1 and Access B2) and 1 services/control network, which are connected through a common core network. Likewise, the ISP C's network encompasses 2 access networks (Access C1 and Access C2) and 1 services/control network. Also, a ROMEO super-peer that allows for a proper coordination of the media streaming among the ROMEO peers is placed in each ISP (ISP B and ISP C).

As the network is initialized and set to run, every authorized QoS-sensitive service request to a core network is subject to the resource and admission control process defined by the corresponding local RAM. This implies that best-effort services may not be subject to such admission control, depending on local control policies. In particular, a connection request specifies its desired QoS parameters in terms of bandwidth and buffer, together with the related traffic characteristics (e.g., flows IDs, source and destination IP addresses and ports), as being the original QoS requirements of the request, in a NSIS compliant protocol [17]. This process

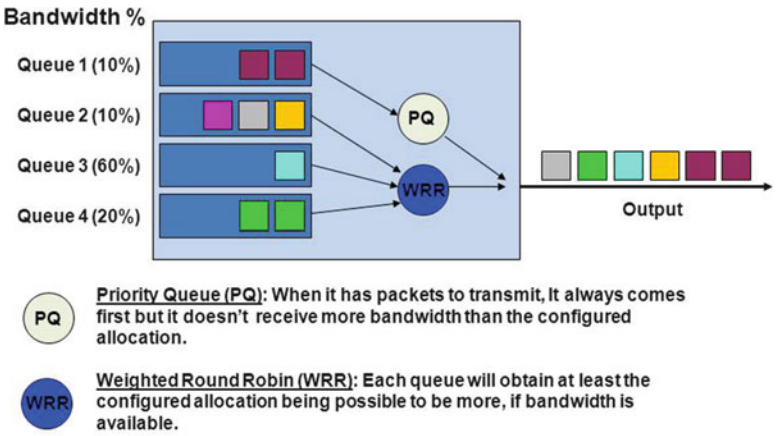


Fig. 10.12 Sample scheduler queuing algorithm

Table 10.1 Traffic parameters for the sample scheduler

Type of traffic	DSCP label	Class	Prior. 802.1p/ Q	Output queue	Throughput	Buffer size
VolP traffic	46 (EF)	5	5	1 (Priority)	10 %	10 %
Voice and video control traffic	26 (AF31)	3	3	2 (WRR)	10 %	10 %
Rounting protocols	48	6	6			
Spanning tree	56	7	7			
Real-time video	34 (AF41)	4	4	3 (WRR)	60 %	26 %
Gold traffic (1^{st})	16	2	2			
Silver tarffic (2^{nd})	8	1	1	4 (WRR)	20 %	54 %
Background (3^{rd})	0 (BE)	0	0			

may be triggered by the ROMEO server, a RAM from neighboring domains in a hierarchical manner, or from a peer to a parent peer or to a super-peer. Thus, a connection scope may be a single network domain (e.g., a connecting child peer and the parent peer(s) are in the same network) or span over several domains, depending on the locations of the initiator and the destination points of the connection. To facilitate the understanding, Fig. 10.14 is used to illustrate the IRACS resource and admission control approach, involving several network domains. In particular, an ISP Y network hosting the ROMEO server and a given peer A in an ISP A's network are connected through a transit ISP T. While RAM is responsible for the QoS control in the core using over-reservation techniques, PCRF controls QoS in the access network based on prioritization and not by reserving resources.

The way the end-to-end QoS-aware multicast trees control is performed to connect peers in a scalable fashion is further detailed in the following:

- First, a joining peer A exchanges appropriate information (e.g., peer's capabilities, requested streams and codecs) with the ROMEO server for AAA. During this

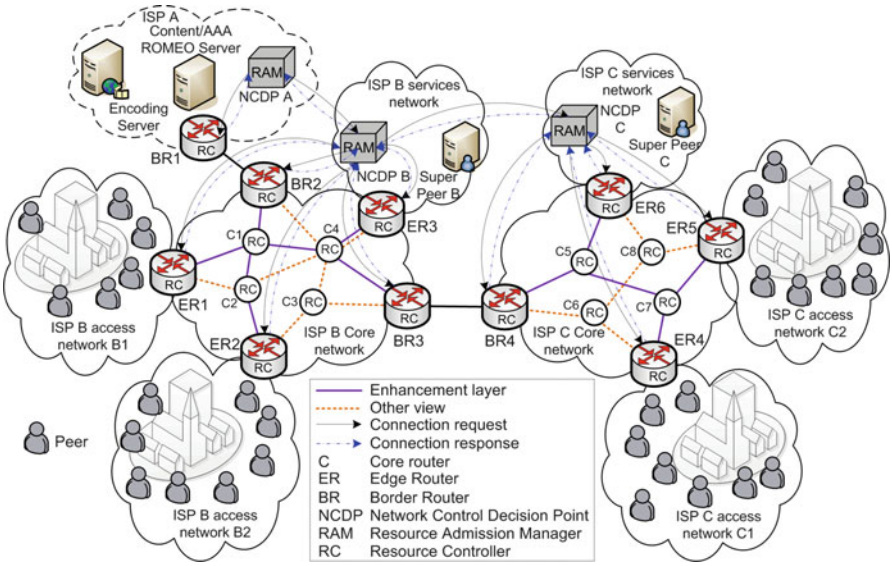


Fig. 10.13 Illustration of large network topology supporting ROMEO

session negotiation phase, the server redirects the peer A to the super-peer A to which it could connect to consume the requested stream. Afterwards, the ROMEO resource and admission process is triggered to perform QoS-aware connectivity between the peers and the server. To better describe the IRACS resource and admission control mechanism over multiple domains as in Fig. 10.14, it is assumed that the requested stream (e.g., a particular enhanced view) by the peer is not yet available at any peer inside the network. This means that the connection must be set up all over the way between the server and the peer.

- Hence, the peer A sends a connection request to the super-peer A along with the desired QoS parameters (e.g., bandwidth and buffer) and the traffic characteristics (e.g., destination peer's ID and traffic flows IDs). Then, the super-peer A triggers the RAM A to obtain QoS-aware trees for connectivity across the core domains. Upon receiving the request, RAM A checks its candidate trees that connects the Edge Router (ER A) to the neighboring transit ISP T on the route towards the server. A SLA should have been preestablished between the involved ISPs. The selected tree must include the super-peer A from which the peer A will receive the content, while the proper BRs (ingresses and egresses in a domain) are obtained based on the inter-domain routing information populated by BGP. In case a candidate tree presents sufficient over-reservation, the RAM A books the best tree among them and forwards the request down to the next RAM T towards the destination without QoS reservation signalling. In case the over-reservation is insufficient, the RAM A invokes the RRS to decide new reservation parameters policies based on the resource information in its local CIB database. If the resource readjustment is successful, the tree is booked and

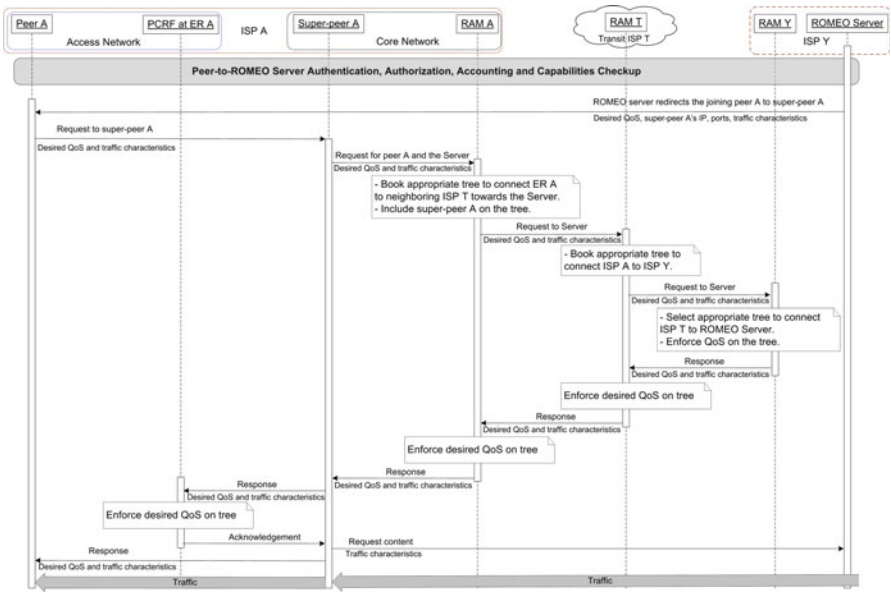


Fig. 10.14 Illustration of IRACS QoS-aware trees control over multiple domains

the message is automatically forwarded to the next RAM T without QoS reservation signalling. This way, RAM A avoids per-flow QoS reservation signalling on trees so as to scale. In case the currently unused network resources are not enough to connect the peer with acceptable QoS, RAM A denies the request and sends a notification back to the super-peer, or it maps the request to a lower QoS CoS according to predefined SLAs between the provider and the customer. This mechanism is repeated at each RAM visited on the route till the message will reach the RAM Y in the ISP domain where resides the ROMEO server. Hence, RAM Y selects the best tree to connect ISP T to the server. Then, it enforces its QoS policies on the tree and sends a response message in the reverse route back to the super-peer.

- Hence, when RAM T intercepts the message and there is sufficient available over-reservation, it simply confirms the desired QoS enforcement in its local database and instructs the related ingress and egress routers to configure the sessions-to-multicast trees mapping so that media packets/chunks can be correctly encapsulated at ingress and decapsulated at egress points. In this case, RAM T forwards the message upstream to RAM A without issuing QoS reservation signalling messages. Otherwise, the RAM T must signal the RCs on the selected tree to readjust the reservations according to the local control policies. After that, RAM A enforces the QoS on its tree and sends the message to the super-peer A that forwards it to the edge node (ER A). At the edge node and through the Network API, the PCRF enforces a higher priority within the PCEF (access domain) to the ROMEO traffic flows over other services and

acknowledges the super-peer. This way, end-to-end QoS-enabled trees are setup and the super-peer sends a response message to the peer A and asks the server to start streaming the media which will enjoy the QoS destined to them in the P2P overlay.

As a result, IRACS deploys a fast and scalable resource and admission control, which efficiently prevents per-flow QoS reservation signalling messages. Moreover, it provides support for service and network convergence by guaranteeing that each admitted traffic flow receives the contracted QoS, without excessive control overhead.

10.5 Summary

This chapter proposes a cross-layer networking control architecture that allows a flexible use of the Internet resource over-provisioning to ease rapid advances in P2P communications. In particular, the approach assists end users' devices to synchronize content delivered through hybrid distribution paths (such as DVB-T2 and the IP networks) in a cost-effective and scalable manner, that is, without high complexity and buffering requirements. In other words, the proposed architecture assures that each connected session will receive a minimum guarantee of its desired QoS in terms of bandwidth and delay through P2P overlay system in support for improved Quality of Experience. The use of efficient resource over-provisioning ensures scalability since the control states and signalling overhead, and therefore the related consumption of CPU, energy, and memory can be significantly reduced. In addition, the dynamic P2P flows (ROMEO flows) prioritization is achieved within the access network thanks to the implementation of the 3GPP PCC architecture that allows P2P services to request flow prioritizations on the fly when the P2P service is going to be consumed. This approach enables to deliver the P2P service with a minimum guarantee having a higher priority over other services consumed by the customer using the same access network.

References

1. ITU-T Recommendation Y (2001) General overview of NGN, Dec 2004
2. http://www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoSIntro.html. Accessed March 2013
3. Blake S, Black D, Carlson M, Davies E, Wang Z, Weiss W (1998) An architecture for differentiated services. IETF RFC 2475
4. ROMEO—Remote Collaborative Real-Time Multimedia Experience over the future internet, FP7 collaborative project. url: http://www.ict-romeo.eu/. Accessed March 2013
5. Demers A, Keshav S, Shenker S (1989) Analysis and simulation of a fair queueing algorithm. ACM SIGCOMM'89 19:1–12
6. Golestani SJ (1994) A self-clocked fair queueing scheme for broadband applications. INFOCOM'94, Networking for Global Communications 2: 636–646, Canada
7. Braden R, Clark D, Shenker S (1994) Integrated services in the internet architecture: an overview. IETF RFC 1633, June 1994

8. Braden R, Zhang L, Berson S, Herzog S, Jamin S (1997) Resource Reservation Protocol (RSVP)—Version 1 Functional Specification. IETF RFC 2205, Sept 1997
9. Manner J, Fu X (2005) Analysis of existing quality-of-service signalling protocols. IETF RFC 4094, May 2005
10. Neto A, Cerqueira E, Rissato A, Monteiro E, Mendes P (2007) A resource reservation protocol supporting QoS-aware multicast trees for next generation networks. In: Proceedings 12th IEEE symposium on computers and communications, Aveiro, Portugal, July 2007
11. Baker F, Iturralde C, Le Faucheur F, Davie B (2001) Aggregation of RSVP for IPv4 and IPv6 reservations, IETF RFC 3175, Sept 2001
12. Logota E, Neto A, Sargento S (2010) COR: an efficient class-based resource over-pRovisioning mechanism for future networks. In: IEEE Symposium on Computers and Communications (ISCC), June 2010
13. Neto A, Cerqueira E, Curado M, Monteiro E, Mendes P (2008) Scalable resource provisioning for multi-user communications in next generation networks. Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008
14. Bless R (2002) Dynamic aggregation of reservations for internet services. In: Proceedings 10th International Conference on Telecommunication Systems—Modeling and Analysis (ICTSM'10), vol 1, pp 26–38, Oct 2002
15. Prior R, Sargento S, "Scalable Reservation-Based QoS Architecture - SRBQ" In: Encyclopedia of Internet Technologies and Applications, Freire M, Pereira M (Eds) IGI Global, Hershey, PA, USA, ISBN: 978-1-59140-993-9, pp. 473–482
16. Sofia R (2004) SICAP, a shared-segment inter-domain control aggregation protocol. Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749.016 Lisboa, Portugal, PhD thesis, March 2004
17. Logota E, Sargento S, Neto A (2010) Um Método para Controlo Avançado de Sobre-reservas Baseado em Classes de Serviço e Sistema para a sua Execução (A Method and apparatus for Advanced Class-based Bandwidth Over-reservation Control)", patent N°105305, Sept 2010
18. 3GPP Technical Specification 23.203 (Rel. 12) Policy and Charging Control Architecture, March 2013
19. Katz D (1997) IP router alert option. RFC 2113, Feb 1997
20. Hancock R, Karagiannis G, Loughney J, Van den Bosch S (2005) Next Steps in Signalling (NSIS): framework. IETF RFC 4080, June 2005
21. Rekhter Y, Li T, Hares S (2006) A Border Gateway Protocol 4 (BGP-4). RFC 4271, Jan 2006
22. Manner J, Karagiannis G, McDonald A (2008) NSLP for quality-of-service signaling, draft-ietf-nsis-qos-nslp-16 (work in progress), Feb 2008
23. Vasseur J-P, Ali Z, Sivabalan S (2006) Definition of a Record Route Object (RRO) Node-Id Sub-Object. RFC 4561, June 2006
24. Braden R, Zhang L, Berson S, Herzog S (1997) Resource ReSerVation Protocol (RSVP)—Version 1 Functional Specification, IETF RFC 2205, Sept 1997
25. Moi J (1998) OSPF version 2, IETF RFC 2328, April 1998
26. ITU-T Recommendation E (164) The international public telecommunication numbering plan, Nov 2010
27. 3GPP Technical Specification 23.003 (Rel. 11) Numbering, addressing and identification, March 2013
28. http://www.cisco.com/en/US/technologies/tk543/tk766/technologies_white_paper09186a00800a3e2f.html. August 2005

Chapter 11
Assessing the Quality of Experience of 3DTV and Beyond: Tackling the Multidimensional Sensation

Jing Li, Marcus Barkowsky, and Patrick Le Callet

Abstract Quality of experience in 3D media requires new and innovative concepts for subjective assessment methodologies. Capturing the observer's opinion may be achieved by providing multiple voting scales, such as 2D image quality, depth quantity, and visual comfort. Pooling these different scales to achieve a single quality percept may be performed differently by each human observer. The chapter dives into the complexity of this subject by explaining the QoE concept using 3DTV as an example. It explains the meaning of the different scales, the current approaches to assess each of them, and the individual influence factors related to the voting which affects reproducibility of the obtained results. Methodologies for assessing the overall preference of experience using pair comparisons with a reasonable number of stimuli are provided. The viewers may also create their own attributes for evaluation in the Open Profiling methodology which has been recently adapted for 3DTV. The drawback of all these assessment methods is that they are intrusive in the sense that the assessor needs to concentrate on the task at hand. Medical and psychophysical measurement methods, such as EEG, EOG, EMG, and fMRI, may eliminate this drawback and are introduced with respect to the different QoE influence factors. Their value at this early stage of development is mostly in supporting and partly predicting subjectively perceived and annotated QoE. The chapter closes with a brief review of the most important technical constraints that impact on the capture, transmission, and display of 3DTV signals.

J. Li (✉) • M. Barkowsky • P. Le Callet
LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, Rue Christian Pauc, BP 50609 44306 Nantes, France
e-mail: jing.li2@etu.univ-nantes.fr; Marcus.Barkowsky@univ-nantes.fr; Patrick.LeCallet@univ-nantes.fr

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_11, © Springer Science+Business Media New York 2014

11.1 Definition of 3D QoE

The term "Quality of Experience" (QoE) unites a multitude of meanings. Some of them were attributed to QoE and similar terms such as "Quality of Service" (QoS) in an ambiguous manner. Recently, representatives of more than 20 internationally recognized research institutions discussed this issue within the European Network of Excellence "Qualinet" (COST IC2003). They decided for the following working definition: "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [1].

The need for this definition has been triggered partially by the recent development of 3D video quality assessment methodologies. While it is evident that multimodal services, such as audiovisual services, require multidimensional quality analysis, 3D video quality assessment is a particularly interesting example of a monomodal service stimulating the human's quality perception in a complex manner that may be modeled in a multidimensional approach.

11.1.1 Multidimensional Perceptual Scales for 3D

The multidimensionality in 3D video QoE is explained by the enhanced depth perception due to the stereopsis effect implemented, in most cases, by projecting two different images to each of the two human eyes and thus mimicking the real world situation in a fixed head position. The technical implementation in 3D capture systems and 3D display devices has deficiencies leading to visual annoyances such as visual discomfort sensations or visual fatigue symptoms.

Several models have been proposed to explain the human's integration of the different aspects, an excerpt will be provided here. Seuntiens et al. proposed a combination of perceived depth, binocular image distortion, and visual strain to model viewing experience in the presence of crosstalk in 2005 [4]. Kaptein et al. proposed to enhance the well-known 2D image quality measurement by adding a depth evaluation and the combination would then lead to a notion of naturalness [2]. This model has further been refined by Lambooij et al. towards a two level perceptual process which measures image quality and amount of depth as primary indicators and naturalness and viewing experience as derived, higher level indicators [3]. Chen et al. added visual comfort as primary indicator to the model and noted that two levels of derived perceptual criteria may be appropriate. He positioned naturalness and depth rendering on the second level and visual experience on the third level [5] leading to the pyramidal representation shown in Fig. 11.2.

11.1.1.1 Added Value of Depth

The depth perceived in stereoscopic 3D (S-3D) reconstruction maintains all previously perceived 2D depth cues, such as occlusion, relative size and relative density, height in the visual field, aerial perspective, texture gradients, light and shading, and linear perspective. Most of the current S-3D displays are limited to two views, such as polarized passive or active shutter glasses displays. These displays would then add binocular disparity and eventually the convergence state of the eyes as depth cue. Autostereoscopic displays provide more than two views and may therefore also reconstruct motion parallax to a certain extent. Cutting and Vishton have analyzed the just-noticeable difference of object's depth position [8]. They observed that binocular disparities may offer an important depth position cue at short distances which decreases linearly with log distance. At a viewing distance of about 1.5 m, a typical viewing distance for a 42 in. screen, the depth resolution of the human eye would correspond to about 1.5 cm. Using a visual depth acuity threshold of 20 arcsec, the minimum perceivable depth difference would correspond to about 9.4 cm in the same situation. On an autostereoscopic display, a psychophysical test has shown that the perceived noticeable difference may be in between these values [6].

The disparity distribution as shown to the observers mostly influences the perceived depth quantity effect. The qualitative effect of depth also relates to the reconstruction of the depth volume, in particular the relationship between horizontal and vertical compared to depth extents. Extreme depth compression may lead to cardboard effects or even puppet theater effect. To improve perceived depth quality, a stereoscopic shooting rule was developed to allow for improved reconstruction of S-3D content using two camera models [7].

11.1.1.2 Visual Discomfort

The added binocular depth introduced by S-3D technology may provide viewers not only a totally different and enhanced viewing experience but also visual discomfort and visual fatigue issues. Recently, visual discomfort and visual fatigue gained increasing attention as it impedes viewers' quality of experience significantly, besides, of course, being related to the viewers' health and safety.

Visual discomfort and visual fatigue are two distinct concepts though they are often confused and interchangeably used in literatures. Visual discomfort is a subjective sensation which accompanies with the physiological change, thus can be measured by asking the viewer to report its level. Visual fatigue refers to a decrease in performance of the human vision system produced by a physiological change, which can be objectively measured and diagnosed [9, 10]. Here, we focus on the visual discomfort issues.

Vergence-accommodation conflict is a well-known factor that would induce visual discomfort [11, 12]. When viewing an object by means of a 3D screen, the

eyes will converge to the virtual object which is in front of or behind the screen plane. However, the accommodation has to be performed at the screen depth level, which is unnatural and will not happen in our daily life. The larger this discrepancy between vergence and accommodation gets, the higher the possibility that observers will perceive visual discomfort.

The comfortable viewing zone, which determines a maximum threshold for the vergence-accommodation mismatch that would not induce visual discomfort, was thus investigated and defined in different ways. Yano et al. [13] proposed that depth of field (DOF), which refers to the range of distances in image space within which an image appears in sharp focus, can be used to define the comfortable viewing zone in terms of diopters (D). A value of ± 0.2D is suggested [14, 15]. Another definition on comfortable viewing zone was based on the results of empirical measurements, in which ± 1 arc degree of visual angle was used [16, 17]. If considering the screen parallax, the comfortable viewing zone can be defined by a percentage of the horizontal screen size. For 3D television, values of ± 3 % were suggested [10].

Besides the large disparity magnitude, studies showed that the parallax distribution might introduce visual discomfort as well [18, 19]. To prevent or avoid visual discomfort, the upper part of the screen should be located further away from the viewer with less parallax dispersion, and the entire image should be positioned at the back [20].

Binocular distortions or binocular image asymmetries seriously reduce visual comfort if present to a sufficient extent [21]. Asymmetries can be classified as optics-related errors, filter-related errors, and display-related errors. Optics errors are mainly geometry differences between the left and right images, e.g., shift, rotation, magnification, or reduced resolution. Filter-related errors are mainly photometry differences between the two views, e.g., color, sharpness, contrast, and accommodation. The main source of error induced by display system is crosstalk. Crosstalk produces double contours and is a potential cause of discomfort [22]. A study showed that vertical disparity, crosstalk, and blur are most dominant factors when compared with other binocular factors in visual comfort [21].

Fast motion can induce visual discomfort even if the object is within the comfortable viewing zone [15, 17, 23]. Motion in 3DTV can be classified into planar motion (or lateral motion) and in-depth motion. Planar motion means that the object only moves in a certain depth plane perpendicular to the observer while the disparity does not change temporally. In-depth motion, which is also called motion in depth or z-motion, is defined as object movement towards or away from an observer [24].

Studies showed a consistent conclusion on the influence of in-depth motion velocity on visual discomfort, i.e., visual discomfort increases with the in-depth motion velocity [10, 15, 17, 25]. However, the influence from disparity amplitude (disparity range) and the disparity type (crossed or uncrossed) of in-depth motion on visual discomfort was still under study. In [17], the results showed that disparity amplitude (magnitude) is not a main factor for visual discomfort. However, the same authors in their recent study [10] showed that visual discomfort increases with

the disparity amplitude. Furthermore, the results also showed that crossed in-depth motion would induce significantly more visual discomfort than the uncrossed and mixed conditions (both crossed and uncrossed).

For planar motion video sequences, studies showed that visual discomfort increases with the motion velocity [10, 23, 26, 27]. However, the influences of the disparity on visual discomfort led to different conclusions in these studies. Basically, the conclusions can be classified as two types. One is that the disparity type, i.e., crossed and uncrossed disparity, did not affect the visual discomfort thresholds [26] in which the vergence-accommodation conflict has a significant impact. The other is that the crossed disparity will generate more visual discomfort than the uncrossed disparity [10, 23, 27]. A possible explanation for these two different conclusions might be the position of the background. For the first conclusion derived by [26], the background was positioned at the screen plane. For the latter conclusion derived by [10, 23, 27], the positions of the background were placed at a fixed position behind the screen.

11.2 Subjective Assessment Methodologies for 3DTV

The complexity of perceiving S-3D content as opposed to real-world perception explains the difficulties that naïve observers experience when asked to provide an opinion on the QoE of a particular video sample. On the one hand side, they have limited experience with the new technology, notably as opposed to 2D television and, eventually, multimedia content. On the other hand side, they may need to counterbalance positive and negative effects such as added depth value and visual discomfort.

11.2.1 Observer Context Dependency

An observer participating in a subjective assessment experiment cannot be considered isolated from his previous experience and current status. He bases his internal vote on many influence factors which he then expresses towards the outside world, mostly in the form of a vote on a limited scale. Figure 11.1 lists his external experience on the left, notably situations which he has encountered himself, termed "reality," experience with currently available, often widespread reproduction technology such as 2D television, and new reproduction technologies such as S-3D. He uses his perception towards the goal of analyzing the scene information itself and the perceived artifacts which is the main task that he is asked to perform. However, he also consumes and interprets the perceived visual and – eventually – auditive information leading to a match or mismatch with his experience in reality. Last but not least, he also takes into consideration his overall feeling, notably his health conditions which may be divided into perception intrinsic factors, i.e., those related

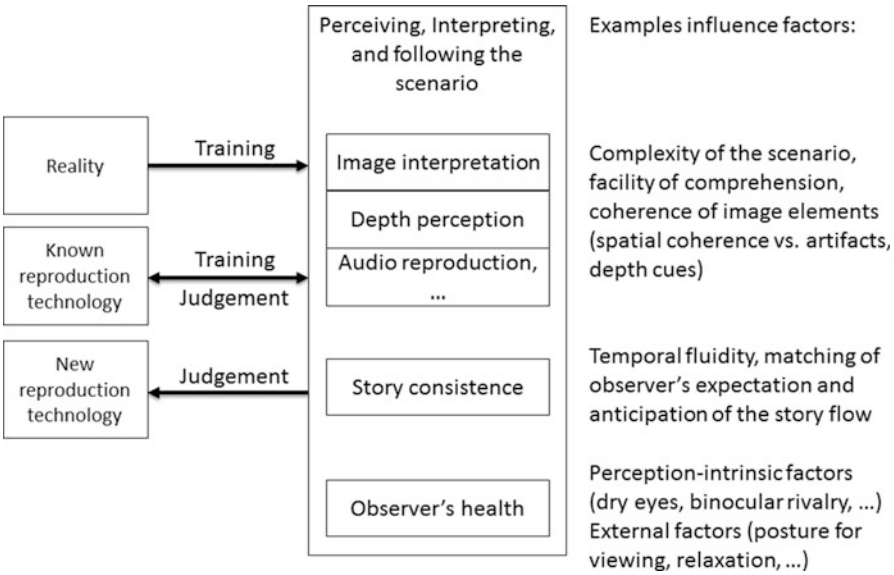


Fig. 11.1 Model excerpt for a human observer in a subjective assessment task

to eyesight, and other health factors which may or may not be related to the task at hand.

An example of this context dependency is related to one of the major decision factors when introducing S-3D services: their advantage over 2D content. From a subjective assessment point of view, the observer's habit to watch 2D content on known reproduction technologies is often misleading their judgment for 3D content shown on the new reproduction technology. Their prejudice may impact in two opposite directions. Often, they judge the 3D content mostly on their trained 2D quality aspects, i.e., perceived coding artifacts, blurring degradations, or reduced resolution, for example, when judging 3D content on a vertically view-interlaced polarized display. On the opposite side, some observers overestimate the sensation of depth as a new and exciting experience as part of the so-called hype effect. Comparing 2D to 3D videos will therefore always be context dependent. Even when introducing both media types into a single subjective experiment, observers will likely change the context from presentation to presentation, therefore, for example, either neglecting or overestimating the added depth value.

11.2.2 Multiscale Assessment Methodologies

A possible solution to expressing the observer's opinion in complex and eventually conflicting situations concerning his internal representations of quality may be to

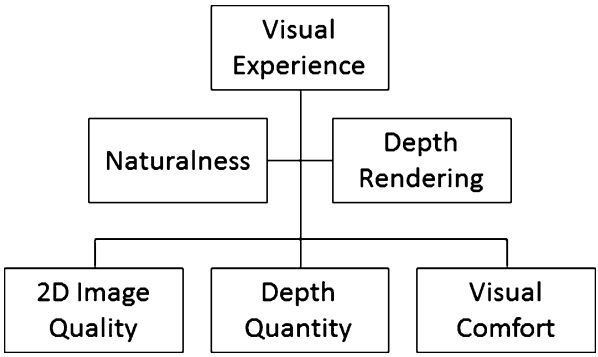


Fig. 11.2 Quality of experience model starting from primary factors on the *bottom* to more complex factors on higher levels

use multiple scales. The observer may judge one aspect such as the perceived image quality independently from other aspects such as the depth quantity or visual discomfort symptoms. These scales have been proposed in Fig. 11.2 as basic 3D quality factors. Several assessment methodologies have been developed to allow for assessing multiple dimensions at once or in separate experiments. Assessing all dimensions in a single experiment facilitates the decorrelation between the scales for the observer, i.e., he decides immediately which effect he assigns to which scale. The advantage of individual experiments with a single scale is reduced experiment duration and focus of the observer on a single quality perception aspect, i.e., he does not need to change his voting context. In most cases, one of the three following standardized methods was used:

- Absolute Category Rating with Hidden Reference (ACR-HR): A single stimulus presentation methodology where the observer votes using a fixed number of attributes per scale, such as the five attributes "excellent," "good," "fair," "poor," and "bad" [32]. High-quality reference sequences are usually included in the experimental setup to allow for calibration of the observer's voting. Each video sequence is presented only once in random order.
- Double Stimulus Continuous Quality Scale (DSCQS): A double stimulus presentation methodology in which the observer watches two different video sequences with one repetition. One of the two video sequences shall be the reference, the other one a degraded version of this reference. He votes for each of the sequences on a semicontinuous integer scale from 0 to 100 which may be annotated with attributes for easier comprehension [33].
- Subjective Assessment Methodology for Video Quality (SAMVIQ): The experiment is ordered by video content. For each of the evaluated video contents, a group of degradations, usually 8–12, is presented in such an interactive interface that the observer may watch each one repeatedly. The reference video sequence is available explicitly and shall be evaluated in a hidden manner among the degraded versions. When the observer has provided his opinion for each scale and each video, he validates his choices and continues with the next content [34].

The International Telecommunication Union—Radiocommunications (ITU-R) has started a new 3D recommendation in 2012 [29]. Besides the three primary

perceptual dimensions "Picture quality," "Depth Quality," and "Visual (Dis) Comfort," it names two additional perceptual dimensions, "Naturalness" and "Sense of Presence." Besides the abovementioned methods ACR and DSCQS, it proposes Pair Comparison and Single Stimulus Continuous Quality Evaluation which is reserved for usage when a single vote for a video sequence is not sufficient but a continuous evaluation is preferred.

All single value voting methods have the drawback that the 3D content display is interrupted after the playback of a single video sequence and a gray frame shall be shown. This distracts the 3D vision on S-3D displays such that the observer requires time at the start of the next sequence before perceiving the 3D effect to its full extent [30]. A solution to this has been proposed by using a continuous playback such as a 3D movie film. Intervals that shall be voted for are marked with overlayed numbers and the observer shall provide a vote for the complete interval [28, 31].

11.2.3 Attribute Selection

Besides choosing the scales for a subjective assessment, the attributes that were used for voting need consideration. When using categories in different languages, important differences may occur, leading to the requirement of aligning the scales from one country to another [36]. It was shown that in many languages the currently employed attributes are not equidistant either and that service acceptance thresholds may vary largely. Assuming that the groups of observers in four different languages would vote for a common average value when judging the same video sequences, a numerical fitting of attributes has been calculated based on the attribute positions for the French scale as used by Zielinski et al. [36]. This led to Fig. 11.3 which shows the experimental results for the 3D experiments with long bars [35] and the results from [36] with shorter bars. While the usual terms "Excellent," "Good," "Fair," "Poor," and "Bad" are used for both "image quality" and "depth quality," the ITU-R has introduced the scale items "Very comfortable," "Comfortable," "Mildly uncomfortable," "Uncomfortable," and "Extremely uncomfortable" for visual discomfort [29]. The drawback of this scale is that the attributes are hard to associate and to distinguish for untrained observers. A typical observer question would be: "How comfortable is 2D viewing on this scale?"

11.2.4 Preference of Experience: Paired Comparison as Ground Truth

In most cases, an overall decision on the quality has to be taken. Multiscale experiments only evaluate a particular quality aspect but the combination of the aspects may be complex. Linear models have been used so far but the stimulus

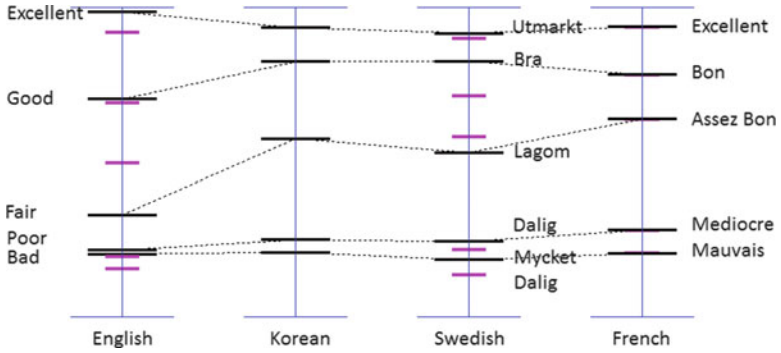


Fig. 11.3 Usage of attributes in four different languages under the assumption that the same MOS value would have been obtained. The *long bars* indicate experimental finding in a 3D QoE experiment, the *shorter bars* the positions published by Zielinksi et al.

degradations were very limited. In addition, the selection of category descriptions for the scales may alter the meaning of the scales in different situations such as viewing contexts or languages, and therefore determining an overall quality remains a challenge. The paired comparison methodology may provide a solution.

The paired comparison methodology is already a standardized subjective video quality assessment method for multimedia applications [32]. The observers compare two video sequences to each other and note their preference. The presentation may be either time parallel, i.e., on two screens, or time sequential, i.e., on one screen.

For m stimuli S_1, S_2, ..., S_m, the test pairs are generated by combining all the possible $N = m(m-1)/2$ combinations S_1S_2, S_1S_3, S_2S_3, etc. When considering the stimulus display order, all the pairs of sequences should be displayed in both possible presentation orders (e.g., S_1S_2, S_2S_1), the number of combinations will raise to $N = m(m-1)$ for one observer. After the presentation of each pair, the observers judge which element in a pair is preferred in the context of the test scenario.

One advantage of the paired comparison method is its simplicity. Observers just need to make a selection for each pair, no judgment scale issue is considered. Another advantage of paired comparison is the enhanced discriminability between similar quality levels compared to the scale rating methodology [37]. Since naive viewers are not used to 3D television and thus have no reference to compare with as in the 2D condition, it might be difficult for the viewers to vote on an absolute psychophysical scale for the stimulus such as "viewing experience" introduced in Sect. 1.1. Thus, the paired comparison method is a possible solution as observers seem to have less difficulties in responding to the question: "which one do you prefer in this pair?" compared to answering "is the quality of this 3D sequence excellent/good/fair/poor/bad?"

The preference vote provided by the viewers may be converted to a scale value indicating his "Preference of Experience" (PoE). PoE represents the preference of

the QoE of the observers in a paired comparison test as the observers only provided their preferences between each two videos rather than an absolute scale value for each video sequence.

However, there is a drawback for the paired comparison method. With the increase of the number of stimuli, the number of comparisons increases exponentially, and thus, it becomes infeasible to conduct the pair comparison experiment. To reduce the number of comparisons, some designs for the pair comparison method are proposed. These designs are introduced briefly here.

11.2.4.1 Balanced Subset Design

Since it is unwieldy to run all pairs in paired comparison method, one possible way is to omit some pairs completely. Dykstra [38] proposed a "balanced subset" method, which means that for certain pairs (S_i, S_j) the comparison numbers n_{ij} is 0 while for all other pairs (S_p, S_q) it is a constant where $n_{pq} = n$. Each of the stimuli has the same frequency of occurrence in the whole experiment which leads to the "balanced" design. Dykstra developed four types of balanced subset design: "Group divisible designs," "Triangular designs," "Square designs," and "Cyclic designs." The "Square design" is briefly repeated here.

Assuming the stimulus number $m = s^2$, the square design (SD) is constructed by placing the m stimuli into a square of size s. Only pairs which are in the same column or row are compared. For example, if there are $m = 9$ stimuli, the indices of the stimulus $S_1, S_2, \ldots, S_9$ could be placed into a square matrix as follow:

1	2	3
4	5	6
7	8	9

In this design, only the pairs among stimuli (S_1, S_4, S_7), (S_2, S_5, S_8), (S_3, S_6, S_9), (S_1, S_2, S_3), (S_4, S_5, S_6), and (S_7, S_8, S_9) are compared. The total number of comparisons is $m(\sqrt{m} - 1)$. The comparison of the number of trials between the full paired comparison (FPC) method and the SD method is shown in Fig. 11.4.

As the SD method only runs part of the pairs, there must be a loss of information. Dykstra gave a definition called "efficiency" to evaluate this method, which showed that this method was highly efficient in predicting the scores of the stimuli. For details, the reader is referred to [38]. However, according to the subjective visual discomfort experiments for 3D videos in [39, 40], the SD method is not robust to observation errors and the influence from the occurrence of other stimuli if the indices of the stimuli were placed into the square matrix randomly. How to arrange the square matrix is an issue which could improve this design to be more robust. Optimized square designs are thus proposed [41], and one of them, called Adaptive SD method is introduced in the following part.

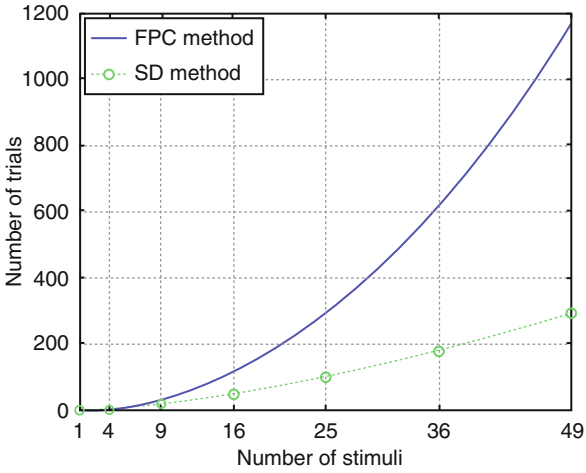


Fig. 11.4 Comparison of the number of trials for full paired comparison (FPC) and square design (SD) method

11.2.4.2 Adaptive Square Design (ASD)

Based on the analysis in [39], comparisons should be concentrated on the pairs with closer quality in the test. Thus, in SD method, the stimuli with similar quality should be arranged in the same column or row to increase the probability to be compared. When the indices of the stimuli are ordered in (descending or ascending) quality as $(d_1, d_2, d_3, \ldots, d_m)$, a possible solution of the arrangement of the square matrix is shown in Fig. 11.5. According to this, an optimized SD method called "Adaptive Square Design" was proposed [40, 41].

The detailed steps about the ASD method are as follows:

- Initialization of the square matrix. If the ordering of the stimuli is unknown, the position of the square matrix could be arranged randomly. Else, the position should be arranged along the spiral as shown in Fig. 11.5 according to the pretest results. Afterwards, run paired comparisons according to the SD rule, i.e., only the stimuli in the same column or row are compared.
- Calculation of the estimated scores. According to current paired comparison results obtained by the previous k observers ($k \geq 1$), calculate the scores and the ranking order.
- Arrangement of the square matrix. According to the order rearrange the square matrix for the $k + 1$ observer, then run paired comparisons for the $k + 1$ observer based on the updated square matrix.
- Repeat step 2 and 3, until certain conditions are satisfied (e.g., all observers finished the test or targeted accuracy on confidence intervals are obtained).

The main difference between the original SD and the ASD method is that the position of each stimulus in the square matrix is updated for each observer. This method has been verified by the subjective visual discomfort experiment in 3DTV and showed robustness for observation errors and other influence factors [40].

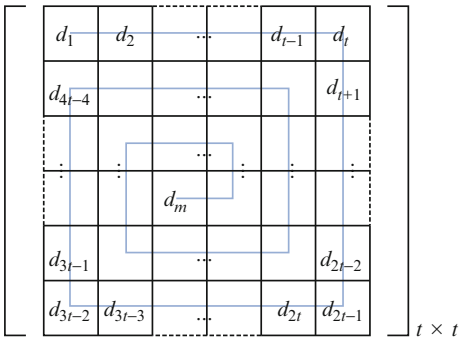


Fig. 11.5 A possible design methodology for optimizing the comparisons performed in adaptive square design (ASD)

11.2.4.3 Sorting Algorithm-Based Design

In [42], based on the idea that comparisons between very distant samples do not provide as accurate an estimate of distance as nearby samples, sorting methods are used for paired comparisons. Firstly, efficient sorting algorithms based on comparing two elements at a time requires M log2 M rather than M^2 comparisons between samples. Secondly, a sorting algorithm should include comparisons between samples of similar quality. This assures that there will be one voting experiment between each closest set of samples, and fewer checks between more distant samples.

An example of using a binary tree sorting method is introduced here. A binary tree can be constructed with the stimuli as the nodes. Each node of the tree is a partitioning element for a left sub-tree and a right sub-tree. The left sub-tree consists of nodes which were judged to be lower in quality, and the right sub-tree consists of nodes which were judged to be higher in quality. During the comparison process, the new stimulus is always compared from the root node. If there is no root node, this stimulus is considered as root. If this stimulus is judged as higher quality, it is then added to the right sub-tree; otherwise, it is added to the left sub-tree recursively. To improve the efficiency of the sorting, a balance process is added after each comparison, which means reconstructing the tree to make it as small as possible and to have as few nodes at the bottom as possible. This method is evaluated by Monte Carlo simulation and showed high efficiency and accuracy in predicting visual quality.

11.2.5 Descriptive Quality Evaluation Methods

Traditional subjective measurement methods have focused on quantitative psychoperceptual evaluations. These methods provide a good basis for examining the relationship between the test stimuli and the sensorial experience. However, as

QoE in 3DTV is related to not only the visual perception of the viewer but also the viewer's expectations, enjoyment, and the evoked emotions from 3D videos, the traditional quantitative methods failed in studying the underlying relationship between the presented stimuli and these individual attributes.

Descriptive quality evaluation approaches, where the participants may define the underlying factors by themselves to classify or differentiate the quality levels, are thus gaining more and more attention recently and have been successfully applied to the user-centered QoE studies on mobile 3D videos [43].

The basic idea of the descriptive quality evaluation methods are to develop the vocabulary (attribute list) used in the subjective test and to evaluate (rate) the test stimuli based on the attributes in the vocabulary. Thus, the contribution of the attributes on the quality of the test stimuli can be analyzed. The implementation methods on the creation of the vocabulary or the evaluation process of the attributes can classify the descriptive quality evaluation approaches into different types. For example, interview-based evaluation methods utilize an individual session per observer called "interview" to create the vocabulary [46]; individual vocabulary profiling methods ask the naïve observers to use their individual vocabulary to evaluate quality, while consensus vocabulary profiling methods require well-trained assessors or experts in the field to develop a consensus vocabulary of quality attributes for quality assessment [44].

Since the traditional quantitative perceptual evaluation methods and the qualitative evaluation methods serve for different aspects of the investigation, a mixed method was also proposed which combines the advantages of both methods and provide a broad interpretation of the results [45].

Open Profiling of Quality (OPQ) [46], a typical mixed method, is briefly introduced here. OPQ is designed to be applicable for naive observers in evaluating the overall quality of the test stimuli. There are in total three steps in the OPQ method. The first step is to estimate the overall quality of the stimuli using the traditional quantitative assessment methods, where the results may serve as a preference ranking of the excellence of all stimuli. The second step is to investigate the underlying characteristics of the stimuli by using qualitative sensory profiling methods, where the individual opinions from naive observers are collected. The third step is a combination of the two results, i.e., analyzing the relationship between quantitative results and qualitative results.

11.2.6 Objective Psychophysical Measurement

Besides the subjective assessment methods, viewers' psychophysical perception and experience in 3D may be predicted by objective psychophysical measurement devices. For example, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) have been used to assess the brain activities which are related to the processing and reactions to the stimuli, e.g., emotion and visual fatigue [50]. Electromyography (EMG) and electrooculography (EOG) are used

to detect the activities related to the eyes, e.g., electrical activity produced by skeletal muscles of the eyes and eye movement, which are considered to provide an indication for visual fatigue.

11.2.6.1 Electroencephalography and Emotive Devices

In clinics, EEG is often used to detect disorders of brain activity or to monitor certain procedures, e.g., the depth of anesthesia. Recently, it has been adopted in psychophysical studies on the relationship between brain activity and 3D QoE because stereoscopic 3D videos would elicit responses from certain brain regions which relate to stereo perception processing, visual discomfort/fatigue, emotion, etc.

The brain waves can be classified into four basic groups according to the frequency band which are gamma (25–100 Hz), beta (14–25 Hz), alpha (8–13 Hz), theta (4–8 Hz), and delta (0.5–4 Hz) bands [47, 48]. Different brain activities can be reflected on different bands. For example, alpha activity is induced by closing the eyes or by relaxation and abolished by thinking or calculating while the gamma band is related to high cognitive processes. These selected responses of brain regions allow for discovering the relationship between the test stimuli and a certain attribute of the QoE. For example, in the study of [49], the authors used an EEG device to compare the brain activity when watching 2D and 3D video sequences. The results showed that the power of the EEG signals in beta frequency was significantly higher when watching 3D contents which might be related to either visual discomfort or visual fatigue.

Another possible application of EEG measurement on 3D QoE is the classification of emotion, which is associated with participant's feelings, thoughts, and behaviors. For example, in [51], the authors used EEG signals to classify happiness and sadness of the participants. The results showed that the emotional states classification is most evident in the gamma frequency band and shows a prediction accuracy of approximately 93 %.

11.2.6.2 Functional Magnetic Resonance Imaging

fMRI is an MRI procedure that measures brain activity by detecting associated changes in blood flow. Compared to EEG, fMRI is more precise in understanding the human brain regions related with the stereoscopic perception due to its high spatial resolutions.

Considerable research efforts have been dedicated to measuring human cortical activity when viewing stereoscopic stimuli [52]. It was discovered that while watching stereoscopic images, the processing and the stereoscopic shape recognition were probably performed in certain regions [53–55]. For example, in [52], the authors used fMRI to test visual fatigue when watching stereoscopic images with disparities of 1, 2, and 3°. The results verified the conclusion that V3A (a cortical

visual area; for more details please refer to [56]) is related to stereoscopic perception as the activation at V3A is much stronger when watching stereoscopic images rather than 2D images. For the stereoscopic images, the results showed that there were strong activities in the frontal eye field (FEF) when watching 3D images with large disparities. This is also supported by a previous EEG study which showed that the areas near the prefrontal cortex (PFC) were related with 3D visual fatigue [57, 58].

11.2.6.3 Electromyography

Usually, EMG is used to analyze the neuromuscular activation of muscles within postural tasks, functional movements, work conditions, and treatment or training regimes. EMG often measures not only at the extremes of the muscle but also along the muscle. As visual fatigue is defined as a decrease in performance of the human vision system produced by a physiological change, it may be desirable to use EMG to detect muscle activities around the eyes and find a relationship with visual fatigue/discomfort.

Nahar et al. [59] studied the EMG response of the orbicularis oculi muscle to "low-level visual stress" conditions, where "low level" means the work conditions in which muscles are activated at a level that can be maintained for a long period of time. The results showed that for conditions in which visual fatigue may stem from eyelid squinting (e.g., refractive error, glare), the power of the EMG response increased with the degree of eyestrain. However, for test conditions without relation to squinting no significant EMG response was measured.

11.2.6.4 Eye Blinking

Eye blinking rate is also considered as an indicator for predicting visual discomfort or visual fatigue. Studies showed that when in relaxed conditions, people would blink more often than in book reading and computer reading tasks [60]. In [61, 62], the results showed that blinking rate was higher in watching 3D video than in 2D. The study of [63] gives the conclusion that eye blinking rate increases with visual fatigue when watching 3D images. For the conditions employing display screens, the blinking frequency was significantly decreased when fatigue was reported (e.g., reading information from the screen for a long time) [64]. In conclusion, eye blinking performs quite differently in different conditions, e.g., in relaxed condition, reading, long-term use of displays, and watching 2D images and 3D images.

Studies in [65] investigated the relationship between eye blinking rate and visual discomfort induced by different types of 3D motion. The results showed that visual discomfort has a linear relationship with eye blinking rate. When watching a still stereoscopic image or 3D video with in-depth motion, the blinking rate increased with the visual discomfort. However, when watching a 3D video with only planar motion, the blinking rate decreased with the visual discomfort.

11.3 Challenges for Measuring and Optimizing QoE in 3D Systems

Most of today's capture, transmission, and reproduction chains were designed for 2D transmission. 3D video transmission schemes are currently under development and may require new guidelines.

11.3.1 Comfort Zone: Shooting 3D Content for Displaying

As opposed to 2D video, capturing and displaying stereoscopic content are not independent. As explained in Sect. 1.1.2, the maximum disparity that can be watched without visual discomfort depends on the display geometry. In addition it is limited for uncrossed disparity to the distance between the human eyes, on average 6.5 cm, such that it does not force the observers to diverge their eyes. The constraint of the comfort zone is related to the coverage of visual angle of a single pixel for a given display device at the designed or preferred viewing distance. The maximum number of depth planes at integer pixel disparity positions that can be reconstructed has been analyzed [14]. As an example, it was shown that in electronic cinema, the maximum disparity should not exceed 14 pixel positions for crossed and 8 pixels for uncrossed disparity. On a 42 in. display, this would correspond to about 25 % of the exploitable comfort zone and therefore lead to a rather poor depth effect. In the opposite case, showing content that was produced for a 42 in. display in an electronic cinema condition, the eyes would be forced to diverge for any uncrossed disparity larger than 8 pixels.

11.3.2 Television Broadcast and Packet Switched Network Video Transmission

Most of today's transmission systems have added 3D as an option rather than being particularly designed for 3D video. The most prominent example is the side-by-side (SBS) format in which most broadcasters transmit 3D content to their clients by reducing the horizontal resolution by half and sending two videos in one single HDTV frame. It is evident that this format impacts not only on the image quality but also on the depth reproduction quality as the disparity information is quantized by a factor of 2 as well.

This backwards compatible format also has the disadvantage that any differences in coding and transmission between the left side of the video frame and the right side result in binocular rivalry conditions when shown on the 3D screen. This is particularly true in case of transmission errors when 2D error

concealment algorithms may alter one part of the video frame. For simulcast coding, it has been shown that observers preferred switching to 2D in this case [66].

An alternative solution is to transmit the pictorial texture information of one or several different camera views and depth map information for each view. At the receiver, the required views are created using depth image-based rendering technologies. The perceived artifacts that this technology introduces differ significantly from previously perceived degradations [67]. In addition, the depth map transmission requires additional bitrate and spatial and temporal subsampling cannot be easily applied [68].

11.3.3 QoE in 3D Display Technology

Current S3D display technologies often require a sampling approach to project the two views into the observer's eyes. Either temporal sampling (active shutter technology) or spatial sampling (passive polarized technology) or color bandwidth sampling (anaglyphic or narrowband color filters) is used. In all cases, some part of the light intensity destined to one eye leaks into the view that is reserved for the other eye. Measurement of this technical parameter called "crosstalk" is a challenging task and cannot be easily compared in between displays. Tourancheau et al. showed differences even on a single model of a particular active display [69]. Recently, the Society for Information Display (SID) finished a recommendation on the measurement methodology for 3D displays including the measurement of crosstalk on currently available display technologies [70].

The crosstalk effect has been studied by Seuntiëns with respect to the influence on overall QoE [4]. An important influence factor is the camera distance and therefore the amount of disparity that accounts for a difference between the two views. An objective model based on subjective data was developed by Xing et al. [71].

11.3.4 Assessment Environment: Lab or Living Room

Generally, the laboratory viewing environment is intended to provide critical conditions to check systems while the home viewing environment is intended to provide a means to evaluate quality at the consumer side of the TV chain. In traditional 2D image/video quality assessment, the influence of the viewing environment is expected to have a significant influence on the observers' results which led to the corresponding recommendations, e.g., ITU-R BT.500 and ITU-R BT.2022. However, the influence of the viewing environment, such as the overall room illumination, the background illumination, and optimal viewing distance for

line interlaced passive displays on 3D QoE, is still under study. Recently, the Video Quality Experts Group (VQEG) started a reevaluation of the viewing conditions in preparation of new recommendations for 3DTV in standardization organizations such as ITU and EBU.

In a similar situation, studies in [72] investigated the influence of the observer's environment on a multimodal, audiovisual subjective test in an international collaboration. Ten different environments were tested in six labs ranging from highly controlled environment to disturbing cafeteria environment. The test results showed that the lighting, background noise, wall color, and objects on the wall were not significant factors or at least were inferior to the influence of the intersubject difference.

This result is similar to the conclusions drawn for a 3D QoE experiment conducted in one lab with exactly the same experimental setup (e.g., displays, participants, viewing distance, stimuli) except for the viewing environment [73]. The experiments were conducted by using the pair comparison method to evaluate the overall QoE of the participants. Two test viewing environments were considered, one was a controlled lab environment which meets the ITU recommendation requirements and the other was a living room environment. According to the statistical analysis, there was no significant influence from test environments.

11.4 Conclusions

Quality of experience symbolizes today a larger variety of perceptual inputs than researchers used to tackle within the 2D television broadcast context for many decades. This chapter aimed at providing an overview of recent developments in terms of 3D QoE measurement. It started from the explanation of the dimensions that stereoscopic 3D viewing may imply. Various methods of assessing the observer's opinion on QoE were provided both in terms of assessment methodologies that use multiple scales as well as single-scale methods, notably paired comparison and the methods which allow to conduct subjective assessment with feasible effort. Approaches which allow users to define their own rating scales were reviewed. Innovative concepts using psychophysical, medical, and emotive measurements in the prediction of QoE were summarized. Last but not least, main impact factors of current technology on each of the dimensions on QoE were provided. Going further into the future, one of the most challenging tasks will remain the understanding of the human's way of interaction with various media, such as the immersion experienced when reading an interesting book that cannot be guaranteed in modern 3D electronic cinemas despite its huge technical complexity, effort, and cost.

References

1. Le Callet P, Möller S, Perkis A (2012) Qualinet white paper on definitions of quality of experience (2012), European network on quality of experience in multimedia systems and services (COST Action IC 1003), (Version 1.1) Published online by COST Action IC 2003, http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf
2. Kaptein RG, Kuijsters A, Lambooij M et al (2008) Performance evaluation of 3D-TV systems. In: Image quality and system performance V. Presented at the society of photo-optical instrumentation engineers (SPIE) conference, vol 6808
3. Lambooij M, IJsselsteijn W, Bouwhuis DG, Heynderickx I (2011) Evaluation of stereoscopic images: beyond 2D quality. IEEE Trans Broadcast 57(2):432–444
4. Seuntiens PJH, Meesters LMJ, IJsselsteijn WA (2005) Perceptual attributes of crosstalk in 3D images. Displays 26(4–5):177–183
5. Chen W, Fournier J, Barkowsky M, Le Callet P (2012) Exploration of quality of experience of stereoscopic images: binocular depth. VPQM, Scottsdale, Arizona
6. Barkowsky M, Cousseau R, Le Callet P (2011) Is visual fatigue changing the perceived depth accuracy on an autostereoscopic display? SPIE electronic imaging: stereoscopic displays and applications, vol 7863
7. Chen W, Fournier J, Barkowsky M, Le Callet P (2011) New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone, SPIE electronic imaging: stereoscopic displays and applications, vol 7863
8. Cutting JE, Vishton PM (1995) Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth. In: Epstein W, Rogers S (eds) Handbook of perception and cognition, vol 5, Perception of space and motion. Academic Press, San Diego, pp 69–117
9. Lambooij M, IJsselsteijn W, Fortuin M, Heynderickx I (2009) Visual discomfort and visual fatigue of stereoscopic displays: a review. J Imag Tech 53(3):1–14
10. Tam WJ, Filippo S, Carlos V, Ron R, Namho H (2012) Visual comfort: stereoscopic objects moving in the horizontal and mid-sagittal planes. In: Society of photo-optical instrumentation engineers (SPIE) conference series
11. Hoffman DM, Girshick AR, Akeley K, Banks MS (2008) Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. J Vis 8(3):1–30
12. Kim J, Shibata T, Hoffman DM, Banks MS (2011) Assessing vergence-accommodation conflict as a source of discomfort in stereo displays. J Vis 11(11):324
13. Yano S, Ide S, Mitsuhashi T, Thwaites H (2002) A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. Displays 23(4):191–201
14. Chen W, Fournier J, Barkowsky M, Le Callet P (2010) New requirements of subjective video quality assessment methodologies for 3DTV, fifth international workshop on video processing and quality metrics (VPQM), Scottsdale
15. Yano S, Emoto M, Mitsuhashi T (2004) Two factors in visual fatigue caused by stereoscopic HDTV images. Displays 25(4):141–150
16. Kuze J, Ukai K (2008) Subjective evaluation of visual fatigue caused by motion images. Displays 29(2):159–166
17. Speranza F, Tam W, Renaud R, Hur N (2006) Effect of disparity and motion on visual comfort of stereoscopic images. Proc SPIE 6055:94–103
18. Ide S, Yamanoue H, Okui M, Okano F, Bitou M, Terashima N (2002) Parallax distribution for ease of viewing in stereoscopic HDTV. In: SPIE proceedings, vol 4660, pp 38–45
19. Nojiri Y, Yamanoue H, Hanazato A, Okano F (2003) Measurement of parallax distribution, and its application to the analysis of visual comfort for stereoscopic HDTV. In: Proceedings of SPIE, vol 5006, pp 195–205

20. Nojiri Y, Yamanoue H, Ide S, Yano S, Okana F (2006) Parallax distribution and visual comfort on stereoscopic HDTV. In: Proceedings of IBC, pp 373–380
21. Kooi FL, Toet A (2004) Visual comfort of binocular and 3D displays. Displays 25(2–3): 99–108
22. Pastoor S (1995) Human factors of 3D imaging: results of recent research at Heinrich-Hertz-Institut Berlin. In: Proceedings of IDW, pp 69–72
23. Li J, Barkowsky M, Wang J, Le Callet P (2011) Study on visual discomfort induced by stimulus movement at fixed depth on stereoscopic displays using shutter glasses, digital signal processing (DSP), 17th international conference on, IEEE
24. Harris JM, Mckee SP, Watamaniuk S (1998) Visual search for motion-in-depth: stereomotion does not 'pop out' from disparity noise. Nat Neurosci 1(2):165–168
25. Cho S-H, Kang H-B (2012) An assessment of visual discomfort caused by motion-in-depth in stereoscopic 3D video. In: Proceedings of the British machine vision conference, pp 1–10
26. Lee S-I, Jung YJ, Sohn H, Ro YM, Park HW (2011) Visual discomfort induced by fast salient object motion in stereoscopic video, SPIE stereoscopic displays and applications XXII, vol 7863
27. Li J, Barkowsky M, Le Callet P (2011) The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos. In: Quality of multimedia experience (QoMEX), 2011 third international workshop on, pp 155–160
28. Gutierrez J, Perez P, Jaureguizar F, Cabrera J, Garcia N (2011) Subjective assessment of the impact of transmission errors in 3DTV compared to HDTV. In: 3DTV conference: the true vision—capture, transmission and display of 3D video (3DTV-CON), pp 1–4
29. International Telecommunication Union—Radiocommunication Sector (2012) Recommendation ITU-R BT.2021: subjective methods for the assessment of stereoscopic 3DTV systems. ITU-R broadcasting service
30. Stelmach LB, Tam WJ (1998) Display duration and stereoscopic depth discrimination. Can J Exp Psychol 52(1):56–61
31. Gutierrez J, Pérez P, Jaureguizar F, Cabrera J, Garcia N (2012) Validation of a novel approach to subjective quality evaluation of conventional and 3D broadcasted video services. In: Quality of multimedia experience (QoMEX), 2012 fourth international workshop on, pp 230–235
32. ITU-T Study Group 12 (1997) ITU-T P.910 Subjective video quality assessment methods for multimedia applications www.itu.ch
33. Question ITU-R 211/11 (1974) ITU-R BT.500-13 Methodology for the subjective assessment of the quality of television pictures www.itu.ch
34. Question ITU-R 81/6 (2003) SAMVIQ subjective assessment methodology for video quality www.itu.ch
35. Barkowsky M, Li J, Han T, et al (2013) Towards standardized 3DTV QoE assessment: cross-lab study on display technology and viewing environment parameters, SPIE electronic imaging stereoscopic displays and applications, vol 8648
36. Zielinski S, Rumsey F, Bech S (2008) On some biases encountered in modern audio quality listening tests—a review. J AES 56(6):427–451
37. Lee J-S, Goldmann L, Ebrahimi T (2012) Paired comparison-based subjective quality assessment of stereoscopic images, multimedia tools and applications, pp 1–18 http://www.springer.com/computer/information+systems+and+applications/journal/11042
38. Dykstra O (1960) Rank analysis of incomplete block designs: a method of paired comparisons employing unequal repetitions on pairs. Biometrics 16(2):176–188
39. Li J, Barkowsky M, Le Callet P (2012) Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. In: Image processing (ICIP), 19th IEEE international conference on, pp 629–632
40. Li J, Barkowsky M, Le Callet P (2013) Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In: IS&T/SPIE electronic imaging, pp 86481V-86481V-12
41. Li J, Barkowsky M, Le Callet P (2013) Subjective assessment methodology for preference of experience in 3DTV. In: 11th IEEE IVMSP workshop: 3D image/video technologies and applications. http://www.ivmsp2013.org/

42. Silverstein DA, Farrell JE (1998) Quantifying perceptual image quality. In: Proceedings of ST&T's image processing, image quality, image capture, systems conference, pp 242–246
43. Strohmeier D, Jumisko-Pyykkö S, Kunze K, Bici MO (2011) The extended-OPQ method for user-centered quality of experience evaluation: a study for mobile 3D video broadcasting over DVB-H. EURASIP J Image Video Process 2011. http://jivp.eurasipjournals.com/content/2011/1/538294
44. Kunze K, Strohmeier D, Jumisko-Pyykkö S (2011) Comparison of two mixed methods approaches for multimodal quality evaluations: open profiling of quality and conventional profiling. In: Quality of multimedia experience (QoMEX), third international workshop on, pp 137–142
45. Tashakkori A, Teddlie C (2008) Quality of inferences in mixed methods research: calling for an integrative framework. In: Advances in mixed methods research, pp 101–119
46. Strohmeier D, Jumisko-Pyykkö S, Kunze K (2010) Open profiling of quality: a mixed method approach to understanding multimodal quality perception. Adv Multimed 2010:1–28
47. Niedermeyer E, Da Silva FL (2005) Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins, Philadelphia
48. Teplan M (2002) Fundamentals of EEG measurement. Meas Sci Rev 2(2):1–11
49. Kim Y-J, Lee EC (2011) EEG based comparative measurement of visual fatigue caused by 2D and 3D displays. In: HCI international 2011–posters' extended abstracts, Springer, pp 289–292
50. Reiter U, De Moor K (2012) Content categorization based on implicit and explicit user feedback: combining self-reports with EEG emotional state analysis. In: Quality of multimedia experience (QoMEX), 2012 fourth international workshop on, pp 266–271
51. Li M, Lu B-L (2009) Emotion classification based on gamma-band EEG. In: Engineering in medicine and biology society, annual international conference of the IEEE, pp 1223–1226
52. Kim D, Jung YJ, Kim E, Ro YM, Park H (2011) Human brain response to visual fatigue caused by stereoscopic depth perception. In: Digital signal processing (DSP), 17th international conference on, pp 1–5
53. Backus BT, Fleet DJ, Parker AJ, Heeger DJ (2001) Human cortical activity correlates with stereoscopic depth perception. J Neurophysiol 86(4):2054–2068
54. Tsao DY, Vanduffel W, Sasaki Y et al (2003) Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. Neuron 39(3):555–568
55. Georgieva S, Peeters R, Kolster H, Todd JT, Orban GA (2009) The processing of three-dimensional shape from disparity in the human brain. J Neurosci 29(3):727–742
56. Tootell RB et al (1997) Functional analysis of V3A and related areas in human visual cortex. J Neurosci 17(18):7060–7078
57. Emoto M, Niida T, Okano F (2005) Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. J Disp Technol 1(2):328–340
58. Li H-C, Seo J, Kham K, Lee S (2008) Measurement of 3D visual fatigue using event-related potential (ERP): 3D oddball paradigm. In: IEEE 3DTV conference: the true vision-capture, transmission and display of 3D video, pp 213–216
59. Nahar NK, Sheedy J, Hayes J, Tai Y-C (2007) Objective measurements of lower-level visual stress. Optom Vis Sci 84(7):620–629
60. Tsubota K, Nakamori K (1993) Dry eyes and video display terminals. New Engl J Med 328 (8):584
61. Lee EC, Heo H, Park KR (2010) The comparative measurements of eyestrain caused by 2D and 3D displays. IEEE Trans Consum Electron 56(3):1677–1683
62. Yu J-H, Lee B-H, Kim D-H (2012) EOG based eye movement measure of visual fatigue caused by 2D and 3D displays, biomedical and health informatics (BHI), 2012 IEEE-EMBS international conference on, pp 305–308
63. Kim D, Choi S, Park S, Sohn K (2011) Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks. In: Digital signal processing (DSP), 17th international conference on, pp 1–6

64. Divjak M, Bischof H (2009) Eye blink based fatigue detection for prevention of computer vision syndrome. In: IAPR conference on machine vision applications, Tokyo
65. Li J, Barkowsky M, Le Callet P (2013) Visual discomfort is not always proportional to eye blinking rate: exploring some effects of planar and in-depth motion on 3DTV QoE. Proceedings of VPQM 2013
66. Wang K, Barkowsky M, Brunnstrom K, Sjostrom M, Cousseau R, Le Callet P (2012) Perceived 3D TV transmission quality assessment: multi-laboratory results using absolute category rating on quality of experience scale. IEEE Trans Broadcast 58(4):544–557
67. Bosc E, Köppel M, Pépion R, Pressigout M, Morin L, Ndjiki-Nya P et al (2011) Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols? In: Proceedings of international conference on image processing, Brussels, Belgium, pp 2597–2600
68. Tam WJ, Speranza F, Vázquez CA, Zhang L (2009) Temporal sub-sampling of depth maps in depth image based rendering of stereoscopic image sequences, SPIE stereoscopic displays and applications XX
69. Tourancheau S, Wang K, Bulat J, Cousseau R, Janowski L, Brunnström K et al (2012) Reproducibility of crosstalk measurements on active glasses 3D LCD displays based on temporal characterization. SPIE electronic imaging stereoscopic displays and applications XXIII, 8288
70. Society for Information Display—International Committee for Display Metrology (2012) Information display measurements standard (IDMS). http://www.sid.org/ICDM.aspx
71. Xing L, You J, Ebrahimi T, Perkis A (2012) Assessment of stereoscopic crosstalk perception. IEEE Trans Multimed 14(2):326–337
72. Pinson M, Janowski L, Pepion R et al (2012) The influence of subjects and environment on audiovisual subjective tests: an international study. IEEE J Sel Top Signal Process 6(6):640–651
73. Li J, Kaller O, De Simone F et al (2013) Cross-lab study on preference of experience in 3DTV: influence from display technology and test environment. In: fifth international workshop on quality of multimedia experience (QoMEX)

Chapter 12
Error Concealment Techniques in Multi-view Video Applications

Carl James Debono and Brian Walter Micallef

Abstract The demand for immersive multimedia experiences is driving researchers and industry to develop solutions that capture, deliver, and display 3D video in an efficient way. The key to the success of this technology is the removal of redundant data through effective video coding and its transmission. One solution that can meet this coding requirement is Multi-view Video Coding (MVC). Besides the removal of spatial and temporal redundancies already used in legacy single-view video transmission, this scheme also considers the redundancies present in between views. Such redundancies exist since the different spatially separated capturing devices are shooting the same view. Removal of the latter significantly enhances the coding efficiency compared to encoding separately each view. This huge reduction in data to be transmitted comes at a price. Practical channels are not error-free, and thus, because of the dependencies generated during encoding, errors in the channel will result in artifacts in the video streams that will propagate in space, time, and views until the dependencies are interrupted. This demands solutions that allow the reconstruction of missing data at the receiver to guarantee a good quality of experience (QoE) which is paramount to the success of 3D video applications. To reduce the occurrence of erroneous data, the information transmitted needs to be protected using error control strategies. These will typically introduce some redundancies that infer knowledge on missing information such that it can be recovered. However, not all the erroneous data can be recovered requesting algorithms that can limit and estimate the missing information. The latter techniques are known as error resilience coding and error concealment methods, respectively, in which the error propagation is restricted and the missing content is estimated from the available information. This chapter will examine the effects of errors experienced during transmission of multi-view video content.

C.J. Debono (✉) • B.W. Micallef
Department of Communications and Computer Engineering,
University of Malta, Msida, Malta
e-mail: c.debono@ieee.org; brian.micallef@ieee.org

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_12, © Springer Science+Business Media New York 2014

This is followed by the application of error resilient and error concealment techniques that introduce enhancements to the quality of the received video content.

12.1 Introduction

Recent advancements in the technologies surrounding the development of three-dimensional (3D) services and applications are pushing towards full immersive multimedia experiences. Video capture systems, information technology tools, broadband network infrastructures, and display technologies all play an important role in the uptake of 3D multimedia services. The capture of 3D video demands the use of multiple cameras to provide the necessary information for the 3D reconstruction. This results in a huge amount of data that needs to be delivered over bandwidth-limited channels. This will add up with the other multimedia content which altogether is expected to exceed 70 % of the total mobile network traffic by 2016 [1]. This suggests that efficient video coding will be a major player to ensure in-time delivery of 3D video data. Displays have also seen improvements in recent years, where stereoscopic displays in conjunction with filtered or synchronized glasses have been available for some time and new auto-stereoscopic displays are becoming available. The latter allow 3D viewing without the need of glasses with a viewing angle proportional to the number of views available at the display. For successful deployment of immersive services, all these components of the technology must ensure a good quality of experience (QoE).

The requirement for 3D video transmission has recently been addressed by the standardization bodies. This has resulted in the Multi-view Video Coding (MVC) extension of the H.264/AVC standard which is the current state-of-the-art CODEC for 3D transmission [2]. Yet, most of the current broadcasts still rely on side-by-side stereo video encoding using H.264/AVC and are sent using MPEG-2 transport streams. MVC allows the transmission of a number of full-resolution camera views and thus can provide a better 3D television (3DTV) experience together with the implementation of free-viewpoint video (FVV) applications [3]. In the latter the user has the ability to select the viewpoint wanted out of many. More recent work has been devoted towards improving quality and coding efficiencies. A coding standard that is expected to replace MVC is the multi-view video plus depth (MVD) [3, 4], which through the depth information allows a better 3D depth impression in 3DTV and better reconstruction of virtual views, using techniques such as depth image-based rendering (DIBR) [5], for smooth navigation in FVV. The MVD is also being studied for the High-Efficiency Video Coding (HEVC) 3D extension [6], where compression gains exceeding 30 % compared to the H.264/AVC counterpart are expected.

The high level of compression is obtained by exploiting redundancies available in the different streams. Single-view block-based video compression techniques rely on the correlation that exists between pixels in a macroblock that is being encoded and its neighborhood and similarly the correlation that exists between

frames in the time domain. The latter implies that if a replacement can be found, only the difference or no data needs to be transmitted, drastically reducing the number of bits needed for transmission. However, this task tends to be complex as a search in a predefined area needs to be done. In multi-view we have another dimension that can be exploited, and this takes the form of inter-view correlation, spanning from the fact that the capturing devices are all focusing on the same scene but from different angles. Hence, the macroblock replacement algorithm can be extended to identify substitutes in other frames. This is the basis of the H.264/MVC and can achieve better coding efficiencies when compared to coding individually each stream for simulcast transmission [7], at the expense of increased complexity.

The data compression allows better use of available bandwidth, making transmission of 3D video viable. However, the dependencies that are introduced during this process affect the quality of the video displayed in case errors occur [8]. Channel errors can be negligible, such as in fiber-optic channels, but can also be significant, such as in wireless channels. As a result of the coding structure of MVC, erroneously received macroblocks will generate artifacts that will propagate in space, time, and in between views. This will drastically reduce the quality of experience of the users, where for 3D videos, it will not only impair annoyance as in single-view video but can also lead to visual fatigue resulting in sickness [9]. Therefore, this problem needs to be mitigated to restore the QoE. Two techniques that can enhance the quality of Multi-View Videos (MVV) are error resilient coding and error concealment methods. In the former case, some redundant data is left with the bitstream together with data interleaving such that erroneous bits can be identified and corrected. However, error control mechanisms, even though they limit the error from propagating, are not always successful in correcting all the errors, and therefore a post-decoding process is required to conceal any remaining erroneous macroblock at the application layer. This data filling exercise is done by interpolating information from correctly received data in the spatial, temporal, and inter-view neighborhood. Thus, a search for the optimal macroblock replacement needs to be done, similar to the way encoding is done. This can highly improve the reconstruction of the 3D data, resulting in a good quality of experience. The process needs to be low in complexity or accelerated via parallel hardware not to delay the displaying of the video.

Even though the standardization of 3D video transmission is young, consumers are already benefiting and consuming content. This is being done through 3D Blu-ray disks, 3DTV broadcasts (mainly over satellite), and the Internet [10]. This technology allows a plethora of services that can be developed, such as educational services, telemedicine applications, remote manipulation of objects, entertainment, and gaming [11]. The growth in this technology can only be guaranteed if adequate quality of experience can be guaranteed through robust error resilient strategies and efficient error concealment.

This chapter will first review the H.264/MVC standard. A study on the effects of channel errors incurred during transmission follows. Some low-complexity error resilience tools will then be discussed. Simple and more advanced error concealment methods will be visited and the quality improvements obtained will be presented. Finally, a conclusion will be drawn with some thoughts over future research directions in this field given.

12.2 Multi-view Video Coding

This section gives an overview of the MVC standard to highlight the structure of the dependencies found in the bitstream and the decoded video streams. The information gained helps in understanding how transmission errors will affect the resulting video, the impact of error resilience, and the reasoning behind the error concealment techniques that can be applied.

Each multi-view video stream has a large amount of spatial and temporal redundancy. A camera is pointing towards a scene which has a number of quasi-homogenous areas. Therefore there is a very good correlation between pixels lying within these areas. This allows a reduction of the data that needs to be transmitted since, at the receiver, macroblocks in the neighborhood can be predicted from known data. This type of coding is known as Intra coding. In a similar way, video frames within a video stream have high rates and thus appear as near stationary if we consider just two consecutive frames. This means that we have a high correlation in the temporal domain that can be exploited to further reduce the data being transmitted. Since there is some movement between the frames, the macroblock in the subsequent frame can be encoded as a motion vector (MV). This is known as Inter frame coding which uses motion compensation techniques to generate the current macroblock from the corresponding macroblock in the previously decoded frame. These Intra and Inter coding schemes are specified in the H.264/AVC standard [2]. Other than these correlations, the MVV have another source of redundancy, which is the dependency that they have in between views, since they are capturing the same scene but from different angles. This dependency is similar to the temporal one, and thus the motion compensation techniques can be extended to replace macroblocks with disparity vectors (DV). This will give a prediction structure as shown in Fig. 12.1. When frames are encoded with no temporal references, to allow random access, these frames are known as anchor frames, such as those at T0 and T8. The additional coding technique provides the basis of the multi-view extension which is defined in annex H of the H.264/AVC standard [2]. Basing the coding technique on this structure produces much less data for transmission than sending all the views as separate H.264/AVC encoded bitstreams.

The entropy encoder takes the sequence of symbols that represent the video and applies a lossless compression scheme to further reduce the size of the bitstream. The two main entropy encoders found in H.264/AVC are the context-adaptive variable length coding (CAVLC) and the context-adaptive binary arithmetic coding (CABAC). The CAVLC applies the variable length coding technique whereby input symbols are mapped on a series of code words that present different lengths using lookup tables, called VLC tables. Typically, frequently occurring symbols will have shorter code words compared to the uncommon ones. The decoder must use the same code word to decode the symbol which in H.264/AVC is available to the CODEC through the use of standard tables defined for generic videos [12, 14]. The CABAC method is based on arithmetic coding that uses a probability model of each syntax element according to its context and adapts the probability estimate

depending on local statistics before coding the sequence. For more information on both schemes, the reader is directed towards [13].

12.3 Channel Errors

The bitstream generated by the encoder needs to be transmitted over some channel. Some channels allow near-error-free transmission, such as fiber-optic channels and satellite links under clear sky conditions. However, other channels, such as the wireless channels, are more prone to errors. This will result in erroneous bits which, in a packet-based transmission scenario, result in the loss of packets. The loss of packets is random and is normally defined by the packet error rate (PER). The errors in the bitstream can be classified as either random errors or burst errors which still have a random occurrence.

The more efficient the compression is, the more susceptible to errors the bitstream becomes. As discussed in the previous section, H.264/MVC employs block-based predictive coding which will cause artifacts generated by the erroneous information to propagate in the direction of the dependency structure generated during encoding. That is, artifacts will propagate in space because of the Intra coding scheme, in time and inter-view because of the Inter coding techniques that use motion estimation and disparity estimation, respectively.

An error in the bitstream can generate an isolated artifact that does not propagate neither in space nor in time, such as an error affecting just one macroblock and possibly its neighborhood but occurs just before an Intra frame (I-frame), an example would be View 0 during T7 in Fig. 12.1. This occurs because the I-frame has no dependency on previous frames. However, in most of the cases, the artifact generated as a result of an error or burst of errors will propagate amplifying the effect of the original error and can drastically reduce the quality of experience of the user. The effect of random errors and burst errors on the *Ballroom* sequence [14]

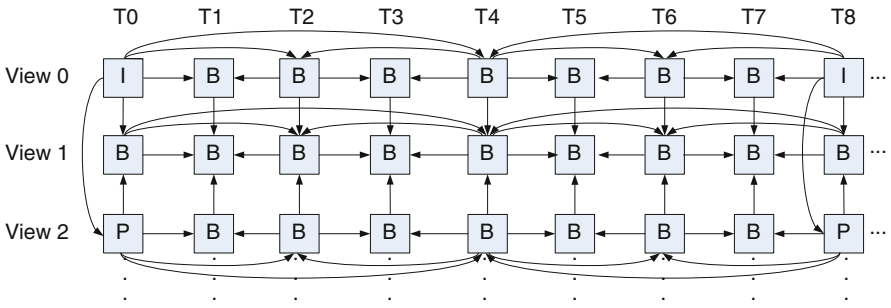


Fig. 12.1 The multi-view prediction structure showing only three views and a group of pictures (GoP) of 8 [12]

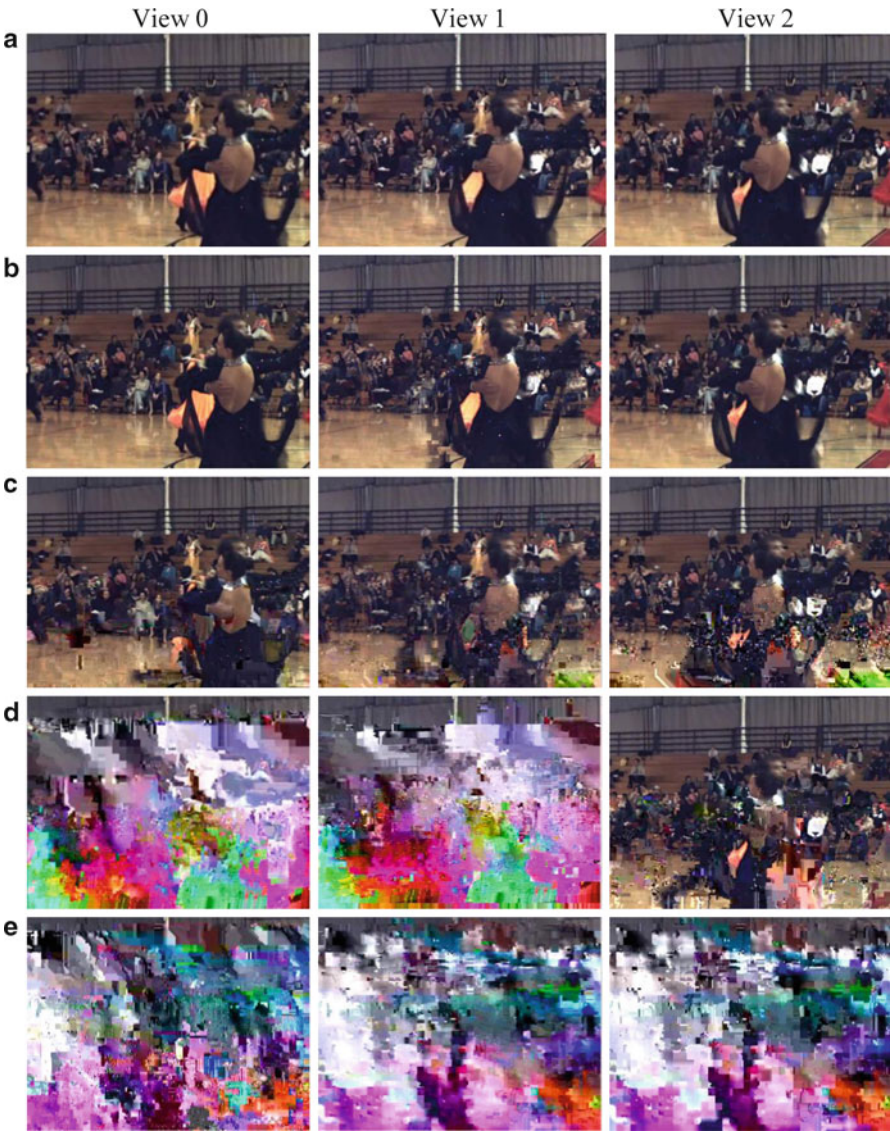


Fig. 12.2 The effects of random errors on frame 71 of the first three views from the *Ballroom* sequence [14], encoded using MVC and the CABAC. (**a**) No errors, (**b**) with a bit error rate of 10^{-6}, (**c**) with a bit error rate of 10^{-5}, (**d**) with a bit error rate of 10^{-4}, and (**e**) with a bit error rate of 10^{-3}

using the CABAC entropy encoding is illustrated in Figs. 12.2 and 12.3, respectively. From these figures it is clear that its effect must be curbed.

The occurrence of an error can result in loss of code word synchronization, and thus the entropy encoder starts interpreting the code words wrongly leading to

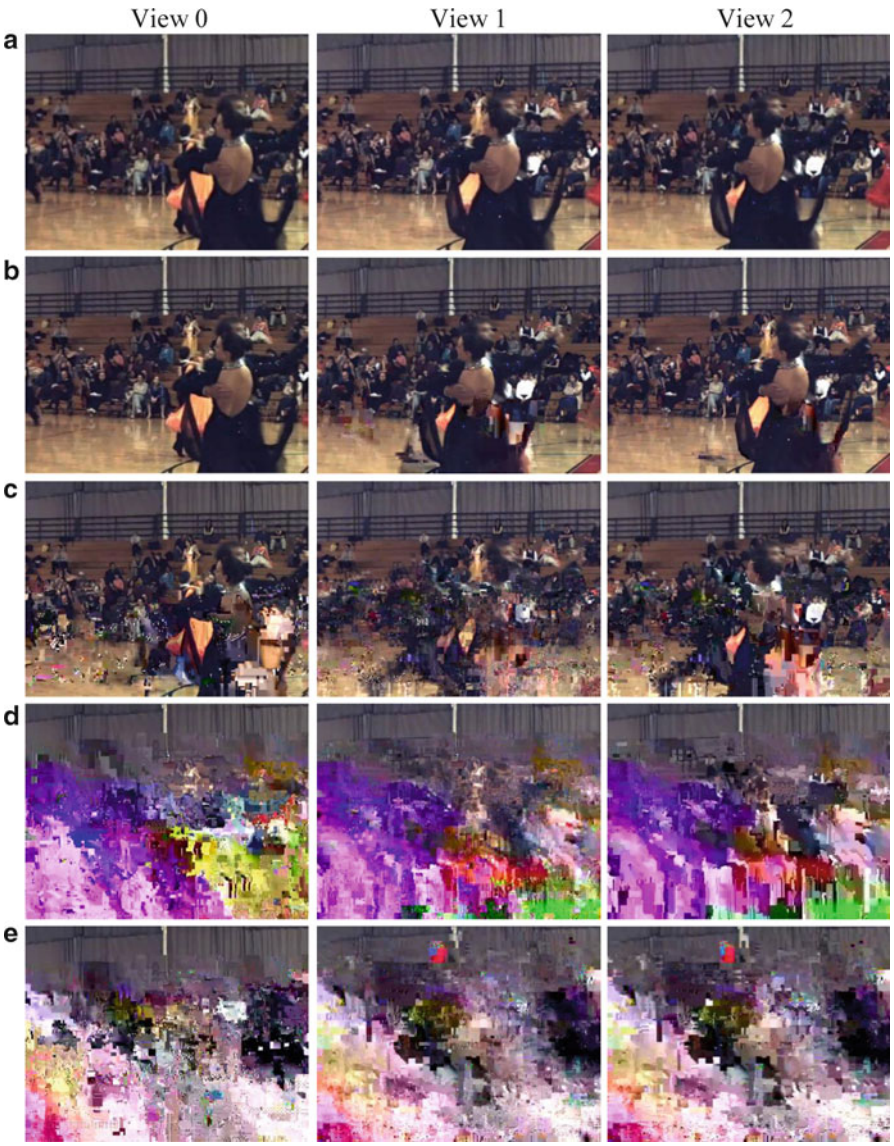


Fig. 12.3 Effect of burst errors, with maximum length of 16 and error probability of 65 % within the burst, on frame 71 of the first three views from the *Ballroom* sequence [14] using CABAC. (**a**) No errors, (**b**) with an error rate of 10^{-6}, (**c**) with an error rate of 10^{-5}, (**d**) with an error rate of 10^{-4}, and (**e**) with an error rate of 10^{-3}

spatial propagation of the error. Such an error will propagate until the next synchronization marker is encountered, which under normal encoding is at the end of the frame. Furthermore, such error propagation can also stem from the loss of coefficient synchronization, where coefficient would be still meaningless without

previous data because of the prediction structure even though code word synchronization is restored.

The propagation of the error on the time axis occurs when motion compensation is applied and the reference macroblock is erroneous. The new macroblock is simply a copy of the erroneous one, just shifted according to the motion vector. This implies that the new frame will have the same artifacts that were in the previous one. The propagation of this error continues until an anchor frame is reached.

A similar error propagation mechanism occurs in the inter-view generated frames where an artifact in a frame of a view will be copied in other views and will also propagate in their temporal dimension. Again the propagation of the error will stop once an anchor frame is reached and a new group of pictures starts.

All the erroneous macroblocks, propagated through the three dependency types, constitute visual effects similar to the ones shown in Figs. 12.2 and 12.3. This analysis shows that error resilience methods are needed to try to limit the impact of errors, correct the errors, and recover the original data. In case these techniques fail, error concealment becomes imperative to minimize the remaining artifacts and recover to some extent the quality of the video.

12.4 Error Resilience Coding

The video data is encapsulated for transmission over the network in Video Coded Layer (VCL) Network Abstraction Layer (NAL) units. These packets are then transported using the MPEG-2 (Motion Picture Experts Group) transport protocols. At the receiver, lower network layer protocols will check the integrity of the each NAL unit, and if the code redundancy check fails, the NAL unit under test is dropped. This means that all the macroblocks encapsulated in such a packet are lost and need to be concealed to regain some quality in the video. The authors in [15] have shown that, in the case of single view, some quality can be gained if the decoder does not drop the corrupted slice and allows decisions on error detection and concealment to be taken at the application layer. The scheme has been shown to be effective also for multi-view video in [8]. Mechanisms to support error detection and bitstream re-synchronization are required to enhance the video data's error resilience.

In the previous section, we have seen that errors can cause the loss of bitstream synchronization, and because of the encoding dependencies, the errors are propagated. Therefore, to limit the proliferation of the errors, fast error detection and action to prevent quality degradation is needed. The decoder, through the syntactic and semantic violation detection tool, is able to detect a syntactic error, such as an illegal code word, or a semantic error, such as a nonexistent reference frame. While providing some basic tool to identify errors, this system leads to a late detection of the error and it provides no error localization method. This means that after the error, synchronization is lost and the error propagates until the next anchor

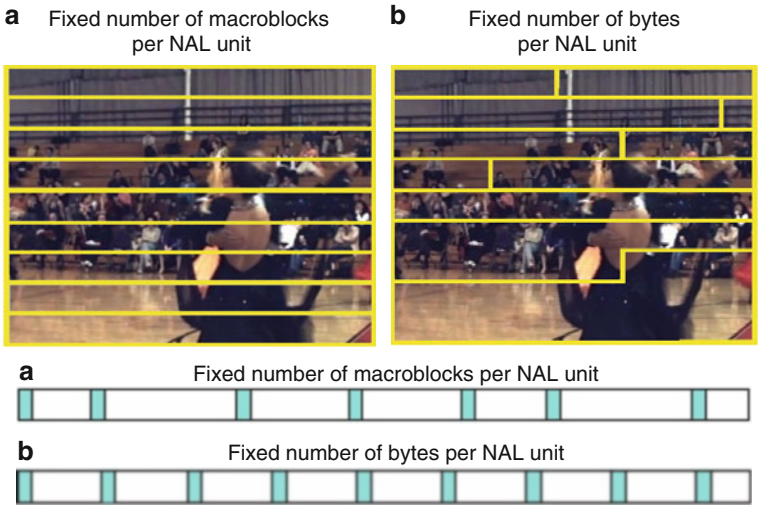


Fig. 12.4 Slice coding with (**a**) a fixed number of macroblocks per NAL unit and (**b**) a fixed number of bytes per NAL unit [16]

frame. Thus, when an error is identified, its spatial propagation should be stopped as soon as possible. This implies that decoder re-synchronization with the bitstream is achieved without dropping too many bits and proceeds with normal decoding. The lost macroblocks will need concealment.

The re-synchronization needed can be accomplished by using the synchronization markers that are available at the beginning of the NAL units. Thus, improvement can be obtained by using smaller predefined NAL unit sizes. This can be implemented using the slice coding tool, leading to slices that contain a group of macroblocks within them, as illustrated in Fig. 12.4. These are encoded without external dependencies such as Intra or motion vector prediction and are decoded independently. This figure shows the difference in having a fixed number of macroblocks per NAL unit and a fixed number of bytes per NAL unit. Smaller cyclic-Intra coded periods lead to more frequent anchor frames, curbing the re-synchronization time and temporal proliferation of the errors.

If the error occurs in the header of the VCL NAL unit or in a NAL unit, this can lead to malfunction of the decoder since information such as frame size, entropy encoder used, and the frame rate parameter can be lost. The amount of bits in these units is small compared to the payload, and therefore in our analysis and evaluation, we will assume that errors will only affect the Raw Byte Sequence Payload (RBSP) of VCL NAL units. Therefore, all important parameters for the decoder are always available and only the encoded data for the macroblocks are lost. This assumption is valid because in practice we can apply unequal protection on the header and thus drastically reduce the probability of this portion of the slices to be in error. An illustration of error propagation when using slice sizes of 200 bytes and CAVLC

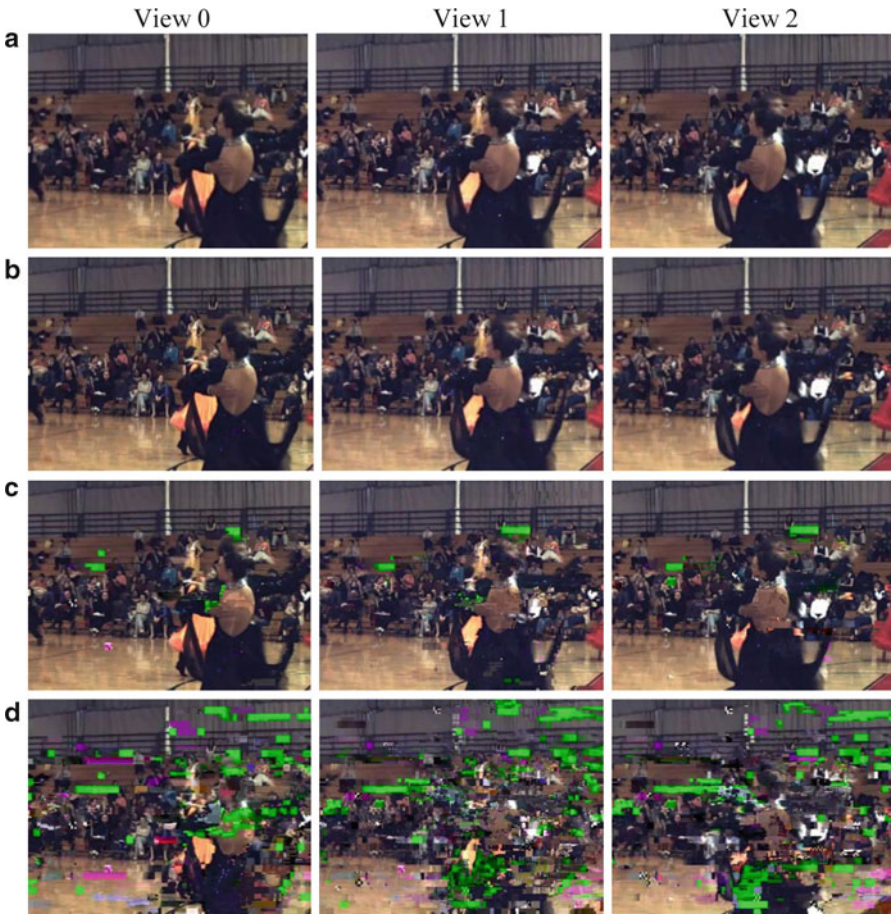


Fig. 12.5 Effect of burst errors, with maximum length of 16 and error probability of 65 % within the burst, on frame 71 of the first three views from the *Ballroom* sequence [14], encoded using CAVLC and slice coding. (**a**) No errors, (**b**) with an error rate of 10^{-5}, (**c**) with an error rate of 10^{-4}, and (**d**) with an error rate of 10^{-3}

entropy encoder is shown in Fig. 12.5. From these results it is clear that even without any concealment, the quality of the video is somewhat improved, since the propagation of the error is now limited.

Making reference to Fig. 12.5, we observe that slice coding has provided some error resilience in that it has curbed the propagation of the errors. It also avoids the loss of complete frames which might be encoded into a single slice if slice size is not limited. Further analyses on the slice size is provided in Fig. 12.6 which represents the visual results when using smaller slice sizes. These results are also captured through average peak signal-to-noise ratio (PSNR) plots in Fig. 12.7.

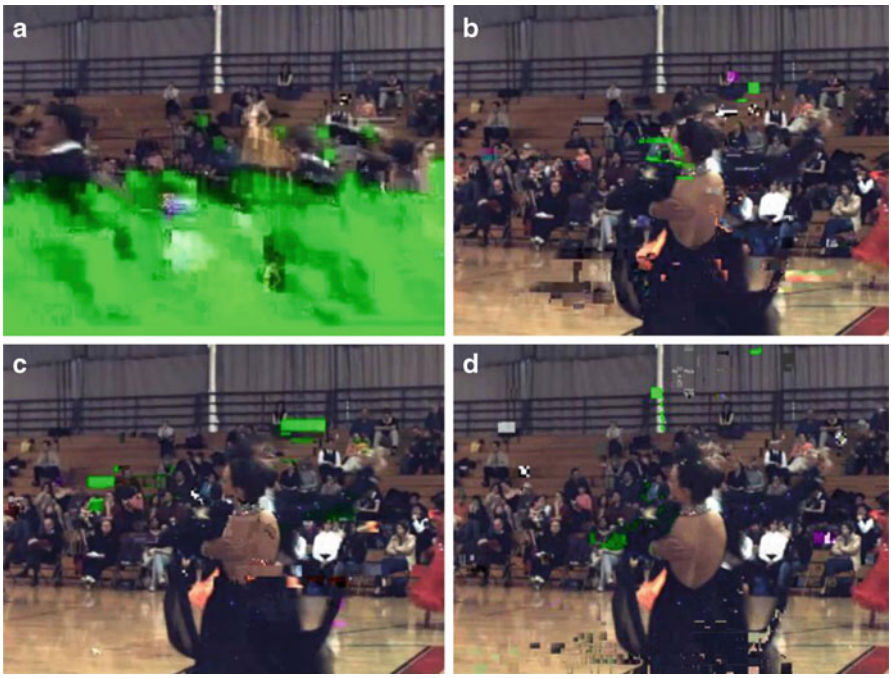


Fig. 12.6 Effect of random errors, with a bit error rate of 1×10^{-4}, on frame 71 of View 2 from the *Ballroom* sequence [14], encoded using CAVLC and slice coding. (**a**) Original, (**b**) slices of 200 bytes, (**c**) slices of 150 bytes, and (**d**) slices of 120 bytes

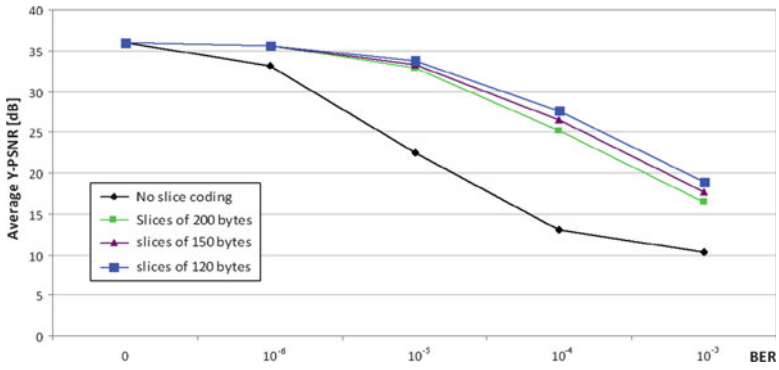


Fig. 12.7 Average PSNR for different slice sizes and bit error rates

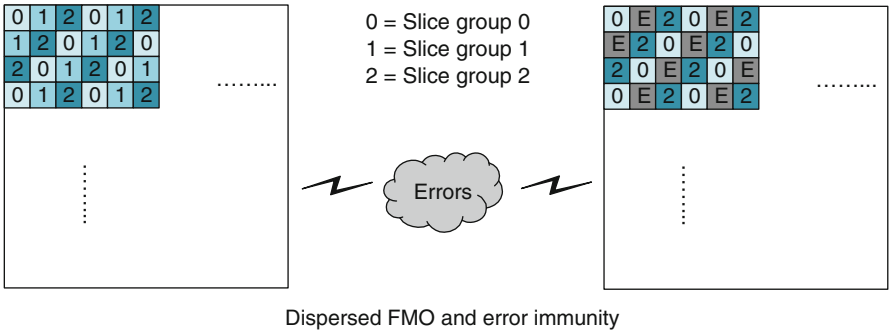


Fig. 12.8 The encapsulation of macroblocks using dispersed FMO with three slice groups

Careful examination of the results reveals that the erroneously decoded macroblocks occur in sequence as packetized during transmission. This will reduce the efficacy of concealment which ideally demands that all neighboring macroblocks are available for efficient missing data interpolation. This suggests that neighboring macroblocks should be encapsulated in different slices, such that loss of a slice will not result in a concentrated artifact but the errors will be dispersed within the frame after its reconstruction. This technique is called flexible macroblock ordering (FMO). We find different scanning patterns for FMO [17]; however, we will be using the dispersed type as it presents good results in such applications. In this FMO type, consecutive macroblocks are placed in different slice groups to ensure that the neighborhood of each macroblock is within a different group as is illustrated in Fig. 12.8 and each slice group is encoded independently, without neighborhood dependencies.

At the decoder, the macroblocks in a slice that is in error will be uniformly scattered on the whole frame. As an example, Fig. 12.8 shows the result of losing a slice containing the macroblocks of slice group 1. From this, it is evident that FMO ensures that we have no accumulation of erroneous macroblocks in a bound region. This means that with the help of slice coding and FMO, there is improved probability that an erroneous macroblock is surrounded by correctly decoded neighbors. Thus we experience a substantial improvement in the quality of the concealment result.

An illustration of the error propagation when applying both slice coding and FMO using slice sizes of 200 bytes and the CAVLC entropy encoder is shown in Fig. 12.9. These results show the effectiveness of the technique which has dispersed the macroblocks in error throughout the whole frame. Further redundancies can be added to provide more data protection such as using forward error-correcting codes. Nonetheless, these are not essential, since the error resilience techniques presented coupled with concealment can give good-quality performances at low-medium PERs [18].

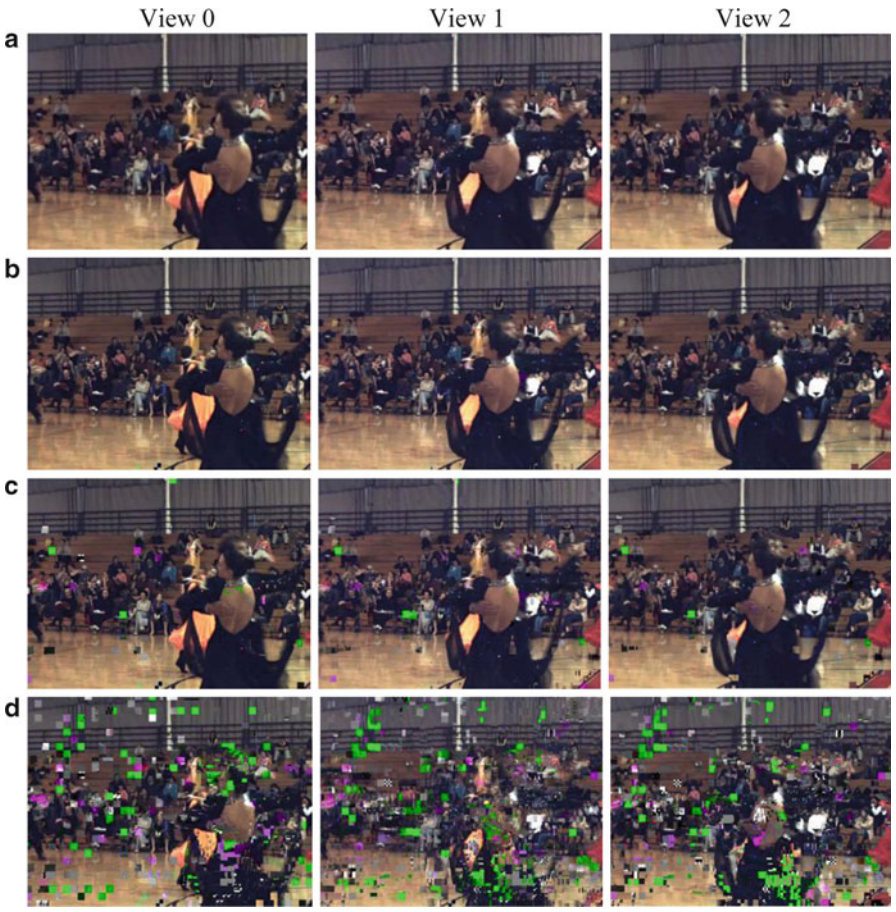


Fig. 12.9 Effect of burst errors, with maximum length of 16 and error probability of 65 % within the burst, on frame 71 of the first three views from the *Ballroom* sequence [14], encoded using CAVLC, slice coding, and flexible macroblock ordering. (**a**) No errors, (**b**) with an error rate of 10^{-5}, (**c**) with an error rate of 10^{-4}, and (**d**) with an error rate of 10^{-3}

12.4.1 Error Concealment

Error concealment is a post-process that is used after the decoder to fill any missing data through interpolation of available data. This can be done because of the correlation that exists between missing macroblocks and their spatial, temporal, and inter-view neighborhood. Techniques available for single-view video can be applied also to multi-view, since similar coding structures are used. Furthermore, the temporal concealment techniques can be extended for inter-view replacements.

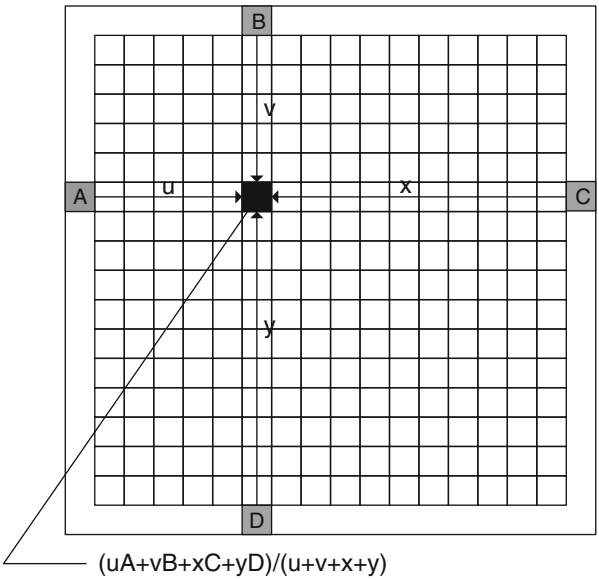


Fig. 12.10 Concept of spatial concealment

An error map containing the locations of all the erroneous and dropped slices is kept during decoding. The order used during macroblock concealment starts from the top and bottom edges of a frame and proceeds towards the center of the lost slices within the map. Similarly, in the horizontal plane, concealment starts at the left and right edges of the image and proceeds towards the center [19]. When an error corrupts macroblocks with an Intra frame, there are no temporal or inter-view references that can be used. In such cases, the missing pixels have to be concealed using a weighted average of the available spatial neighborhood. The interpolation is done using the closest available boundary pixels as illustrated in Fig. 12.10. The weights are evaluated in relation to the inverse distance between the missing pixel element and the reference boundary pixels used. The closer to the missing data the boundary is the better the quality of the concealment. A visual representation of applying spatial concealment to an Intra frame containing erroneous regions is given in Fig. 12.11. In this figure the missing data is shown as black pixels to provide the reader with better understanding of the regions that are being replaced.

For an Inter predicted frame, we can apply temporal error concealment methods. These methods estimate the missing motion vector of a corrupted macroblock using the motion vector information of its spatial neighborhood. A temporal replacement can therefore be found in the previous frame. If the average motion vectors of the neighborhood represent a stationary object or a very slow-moving object, typically representing a movement of less than ¼ of a pixel element, all the missing data is replaced by the collocated macroblocks in the first forward reference frame. If more

Fig. 12.11 Frame 0 of the *Ballroom* sequence [14] transmitted using a 150 byte slice size. (**a**) Original frame, (**b**) received frame assuming a packet error rate of 5 %, and (**c**) the displayed frame after applying spatial concealment

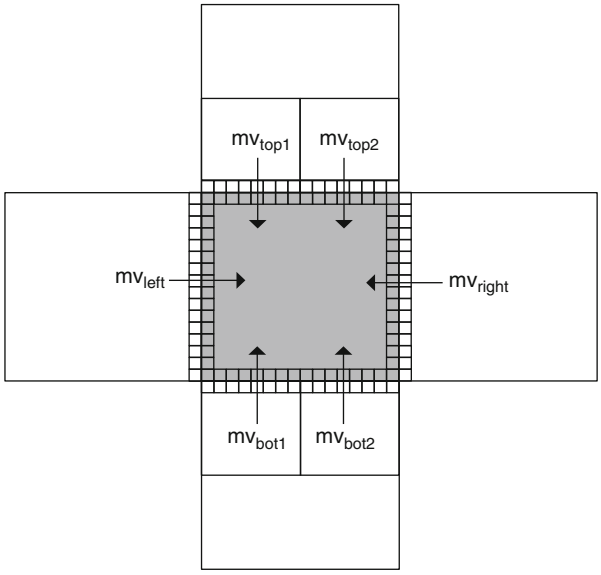


Fig. 12.12 Concept behind temporal concealment

movement is present, a motion vector is selected from the missing macroblock's neighborhood and assigned to this macroblock. Motion compensation of the macroblock is then applied. This represents a good estimate as, statistically, the motion in nearby regions are highly correlated. The concept is represented in Fig. 12.12. An 8×8 luminance block defines the smallest block size that is normally considered for this job. However, macroblocks can be made up of smaller block sizes. In such cases the average of the motion vectors within the 8×8 blocks are used as candidates. We end up with a list of candidate motion vectors for concealment. The selection within this list is then done through the calculation of

the boundary match error. The winning vector is the one that presents the smallest error and is calculated through the boundary matching algorithm (BMA) [21]:

$$d_{\text{sm}} = \underset{\text{dir}\in\{\text{top,bottom,left,right}\}}{\text{minarg}} \left\langle \left(\sum_{j=1}^{16} \left| Y_{\text{ref}}(mv_{\text{dir}})_j - Y_{\text{rec}\ j} \right| \right) / N \right\rangle$$

where d_{sm} represents the side match Luma (Y) distortion, mv_{dir} is the motion vector under test, Y_{rec} is the Luma pixel element at the boundary of the frame being reconstructed, Y_{ref} is the Luma pixel element from the boundary of the motion compensation macroblock, and N is the averaged number of pixel elements. *Top*, *bottom*, *left*, and *right* are the position of the neighborhood macroblock relative to the one being compensated.

This error provides an idea of the smoothness between the image and the motion-compensated macroblock. The zero motion vector must also be placed in the candidate list, since the erroneous data can actually be a SKIP.

In case that the current frame is a B-type frame, we can have more than one motion vector from an 8×8 block. The candidate that will be selected will depend on the existence of these vectors. When only one forward or backward motion vector is available, then selection is trivial. When both are available, then the forward prediction one is used. Note that only correctly decoded neighborhood macroblocks are used for concealment.

If we consider the MVC structure depicted in Fig. 12.1, it is evident that the sequence representing View 0 has no inter-view dependencies. Therefore, this sequence can be viewed as the single-view case and the techniques just described can be applied for concealment of errors within its frames. The Intra coded frames allow random access to the sequence and limit the propagation of errors in the time axis. Weighted average interpolation can be applied to the Intra frames, while motion compensation concealment can be applied to the temporal motion-compensated ones. The second type of encoding sequences in the MVC structure is the single inter-view-predicted view, such as View 2. This has only the anchor frames that are forward predicted from another view. This structure is similar to the temporal prediction case except for the inter-view reference frame that uses the disparity vectors. The non-anchor frames can apply the temporal motion-compensated concealment type. However, the anchor frames do not have temporal references but an inter-view dependency. Therefore, in this case, the disparity vectors of the macroblocks in the neighborhood of an erroneous one can be used for concealment using a technique which is similar to the temporal replacement one [20]. This uses a disparity vector list in which the best match needs to be found. The last type of sequence in the architecture is the inter-view bi-predicted videos, such as View 1. Similar to the anchor frames in the single-view prediction scheme, a lost macroblock can be compensated through disparity vectors from the reference frames. This time we have two reference frames from where to search the best replacement. The non-anchor frames present more options, since the replacement can be obtained using either temporal or inter-view concealment, with the latter

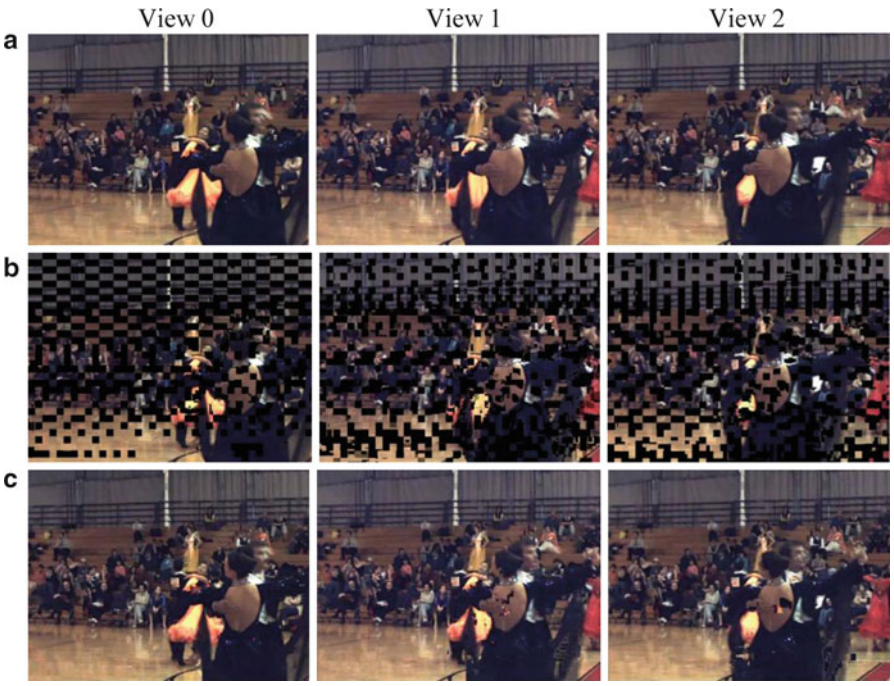


Fig. 12.13 Visual quality improvements on frame 75 of the *Ballroom* sequence [14]. (**a**) Original frame, (**b**) received frame assuming a packet error rate of 10 % on 150 byte slice sizes and using flexible macroblock ordering, and (**c**) the result obtained after applying spatial, temporal, and inter-view concealment

having two views from where to choose. Therefore, the algorithm needs to choose the best reference frame and this depends on the error generated by the motion vectors and disparity vectors being considered.

The results for applying the spatial, temporal, and inter-view concealment techniques to the *Ballroom* sequence [14] can be seen in Figs. 12.13 and 12.14. Figure 12.13 gives a visual interpretation of the effect of these basic methods at a PER of 10 % and using a slice size of 150 bytes. On the other hand, Fig. 12.14 gives the PSNR curves for different PERs, again using a slice size of 150 bytes.

Different objects are found in distinct positions across different views. This implies that they represent different depth values in the scene. An anchor frame is only inter-view predicted, and therefore if a depth discontinuity is encountered, a macroblock is either Intra coded, compensated using a large residual error, or the disparity vector would have lost any correlation with the neighborhood. This occurs because the object is at a different point, is captured from a different angle, or is occluded by other objects in the scene. In these cases, we cannot use the disparity vectors from the neighborhood for concealment, since this would give an erroneous result. Therefore, in such cases, it would be more useful to replace the macroblock using spatial concealment. This suggests that a threshold is used to select between

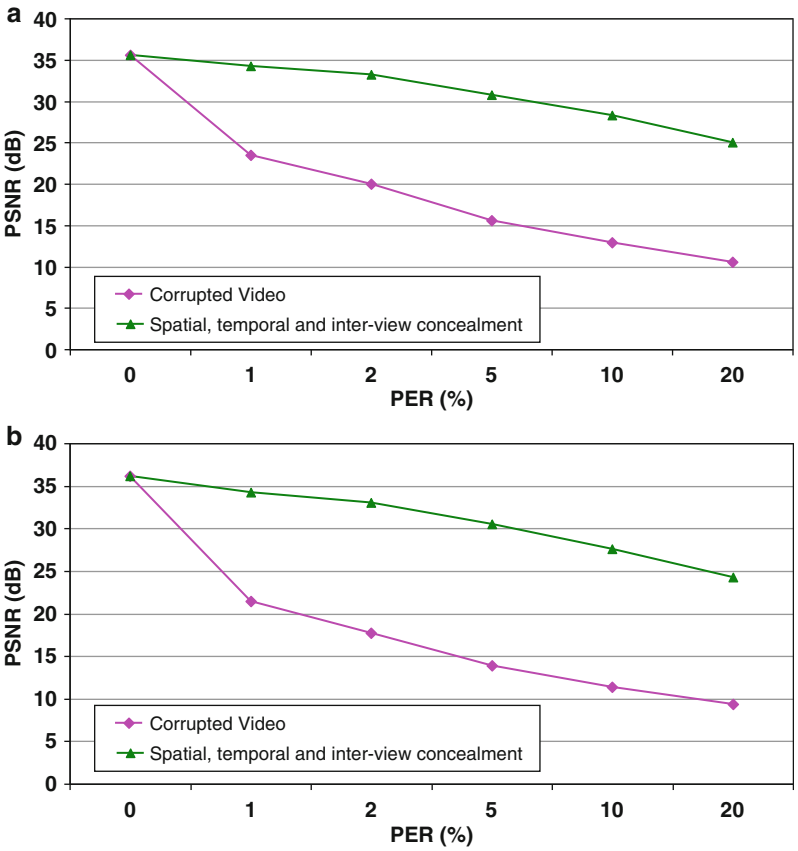


Fig. 12.14 Peak signal-to-noise ratio curves at the different packet error rates. (**a**) Result for View 0 frames and (**b**) average result for inter-view-predicted frames of the *Ballroom* sequence [14]

disparity vector replacement and spatial replacement techniques. The boundary matching distortion error provides us with the measurement of similarity and is used as a means to determine whether the threshold was exceeded. If this error is larger than a predefined threshold, spatial concealment is applied [20].

The visual representation of video data when the disparity vector and spatial concealment threshold system is applied to anchor frames of the *Ballroom* sequence [14] is shown in Fig. 12.15. Furthermore, the gain in performance obtained considering the PSNR curves for different PERs is presented in Fig. 12.16. Note that only View 1 and View 2 are given since anchor frames in View 0 have no inter-view dependencies.

To further enhance the quality of the concealed frames, we can extend the list of possible replacements for erroneous macroblocks. This brings us to a the list containing the following: (1) the zero motion vector, (2) the four nearest neighbor

Fig. 12.15 Visual quality comparison of anchor frame 108 with (**a**) disparity vector replacement and (**b**) disparity vector replacement and spatial concealment method

macroblocks motion vectors for motion compensation, (3) the four nearest neighbor macroblocks disparity vectors for inter-view compensation, (4) the motion and/or disparity vectors of the four corner (top left, bottom left, bottom right, and top right) macroblocks, (5) the average and median of the vectors of these corner macroblocks, (6) the collocated vector from a temporal frame, and (7) the eight neighboring vectors of the collocated vector in the temporal frame. This list presents vectors that have high correlation between the missing data and its spatial and temporal neighbors. For the inter-view-predicted streams, the collocated motion vectors in the Base view (View 0) and the motion vectors of the eight neighboring macroblocks are also considered. To replace the missing macroblock, we need to identify the best replacement vector from the enhanced list. The selection of the motion or disparity vector used to compensate the corrupted macroblock can still be based on the smallest matching distortion parameter measured using the BMA [21]. This algorithm calculates the Sum of Absolute Difference (SAD) between the pixels lying on the edge of the replacing macroblock under test and the outer pixels of the available neighbor macroblocks of the missing block, within the reconstructed frame. The technique is shown in Fig. 12.17. Note that if decoding is being done in real time, only the top and left macroblocks are present and thus only these will be considered in such cases. Simulation results for different PERs are shown in Fig. 12.18, which represents the achieved PSNR on the *Ballroom* sequence [14]. From this figure it is evident that a small gain is achieved through the application of this method.

A variant of the BMA is the outer boundary matching algorithm (OBMA) [22] shown in Fig. 12.19. This algorithm gives the SAD between the two pixel-wide outer boundary of the replacing macroblock under test, coming from its

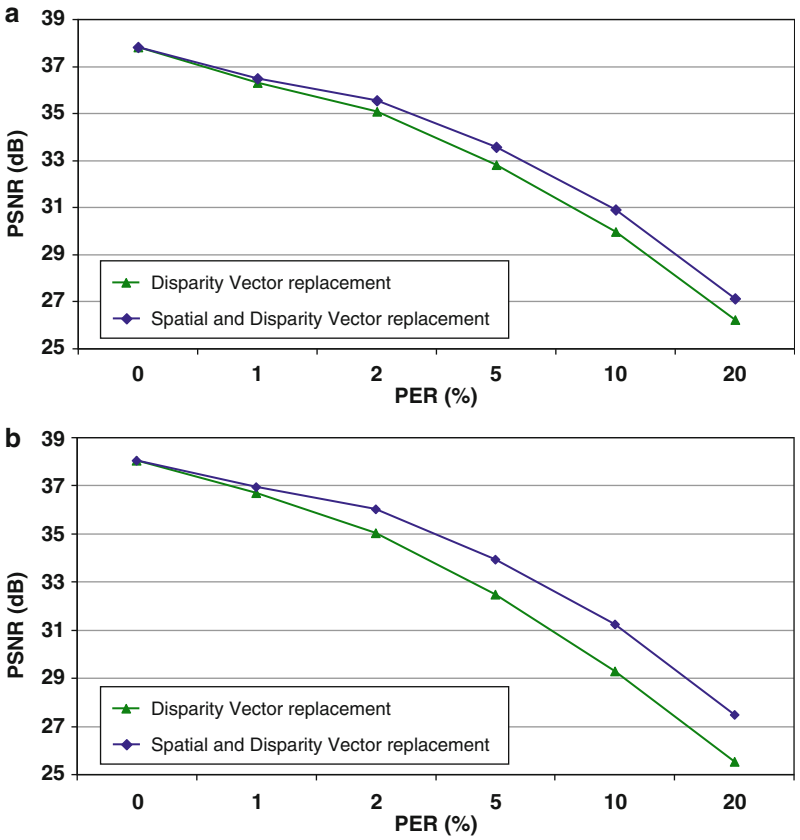


Fig. 12.16 Anchor frame concealment using disparity vector replacement with and without the spatial concealment candidate for (**a**) View 1 and (**b**) View 2

neighborhood in the original frame, and the outer pixels of the available neighbor macroblocks of the missing block. This presents better estimations since it is comparing the pixels located at the same positions in their respective frames. The performance of this algorithm when selecting the appropriate replacement from the enhanced list compared to the BMA is illustrated in Fig. 12.20. These show that better concealment performance is achieved with similar computational complexities [22]. Similar to the case of the boundary matching scenario, if decoding is done in real time, only the top and left macroblocks are available for evaluation.

The selected macroblock for concealment through the above techniques can be further refined through the use of the overlapped block motion compensation (OBMC) method. This algorithm partitions the corrupted macroblock into four 8×8 subblocks. For each of these subblocks, the winning replacement vector and

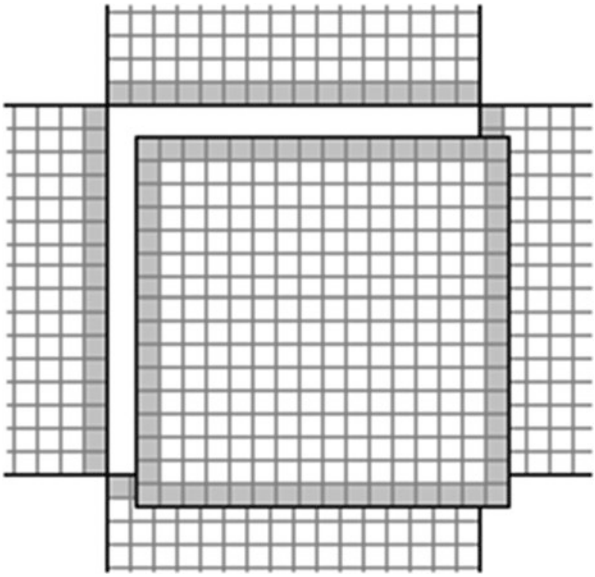


Fig. 12.17 Representation of the inner boundary matching algorithm

the two nearest 8×8 vectors form three motion or disparity compensation subblocks that are then weighted to evaluate the replacement subblock [24]. A subblock compensated with the 8×8 corner vector together with weights that promote the corner compensation macroblock is also added to the subblock replacement weighing, as in [23]. This allows us to select the best motion vector or disparity vector in the list. This solution is an enhancement on the previous algorithms and can be applied after both the BMA and the outer BMA, as shown in Fig. 12.21. The simulation results showing the gain in performance in terms of PSNR are shown in Fig. 12.22. The gains achieved with respect to the previous solutions are evident.

A further addition to the algorithm in Fig. 12.21 is an improvement in the concealment of Intra coded anchor frames. Errors in these frames are typically spatially concealed, yet temporal replacements can still be available [25]. Similarly, disparity compensation can be applied for inter-view anchor frames. This leads to a temporal or inter-view concealment scheme. Therefore, other than the spatially collocated disparity vectors, we can also include the collocated vectors from the temporal frame, the eight vectors in the neighborhood of these vectors, and the zero motion vector of this reference frame, to the candidate list. This expands the list of possible macroblock replacements in the anchor frames, increasing the probability of finding a close match to the missing macroblocks. The results of this technique are given in Fig. 12.23.

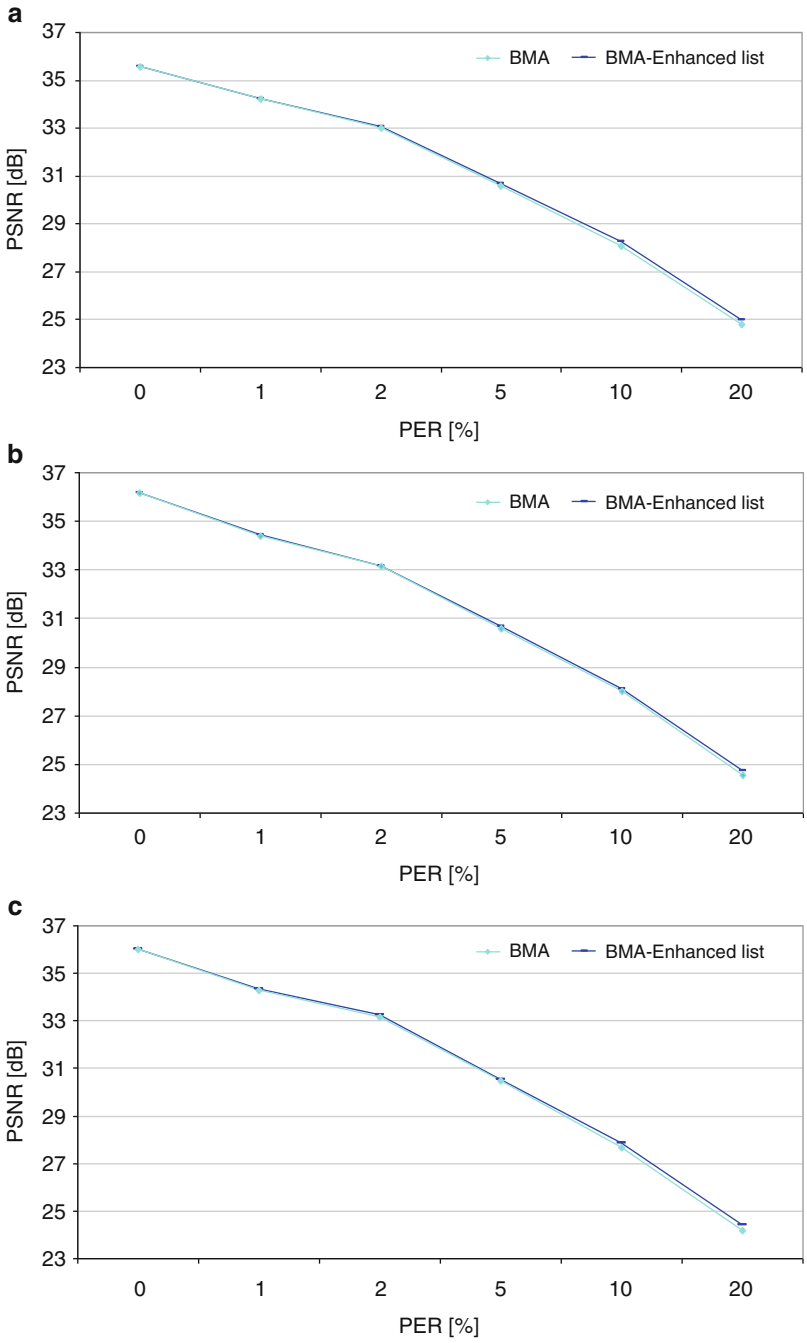


Fig. 12.18 Peak signal-to-noise ratio curves for the *Ballroom* sequence [14] using the spatial, temporal, and disparity vector replacement and the enhanced list with the boundary matching algorithm for (**a**) View 0, (**b**) View 1, and (**c**) View 2

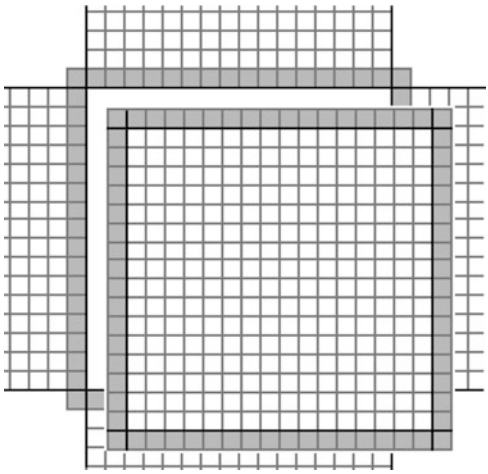


Fig. 12.19 Representation of the outer boundary matching algorithm

The various enhancement methods discussed all lead to some improvement in the quality of the displayed video. An objective comparison, using the PSNR curves, is illustrated in Fig. 12.24. Two sequences [14], *Ballroom* and *Vassar*, are being presented for an appreciation that the algorithms are applicable to different sequences. However, the improvements achieved by these techniques are sequence dependent, since slow movement sequences typically provide more close match replacements than faster sequences.

12.5 Future Directions

The concealment methods discussed all have roots in single-view video and have presented only some extension to exploit the inter-view dependencies. Yet, 3D video transmission is expected to contain the depth information to provide a better feeling of the depth in 3DTV and allow virtual view generation for FVV. This demands the MVD representation which increases the data that needs transmission. The addition of depth information provides more details, in the form of geometric data, about the objects in the video. This geometric information can be exploited to achieve better predictions for macroblock replacements, leading to an increase in performance. This suggests that a different set of concealment tools can be developed specifically for MVD.

The transmission of the MVD has the added challenge that errors can occur in the depth videos. As long as the errors do not affect the edges present in the frame, then the geometric data is preserved and the missing information can be concealed

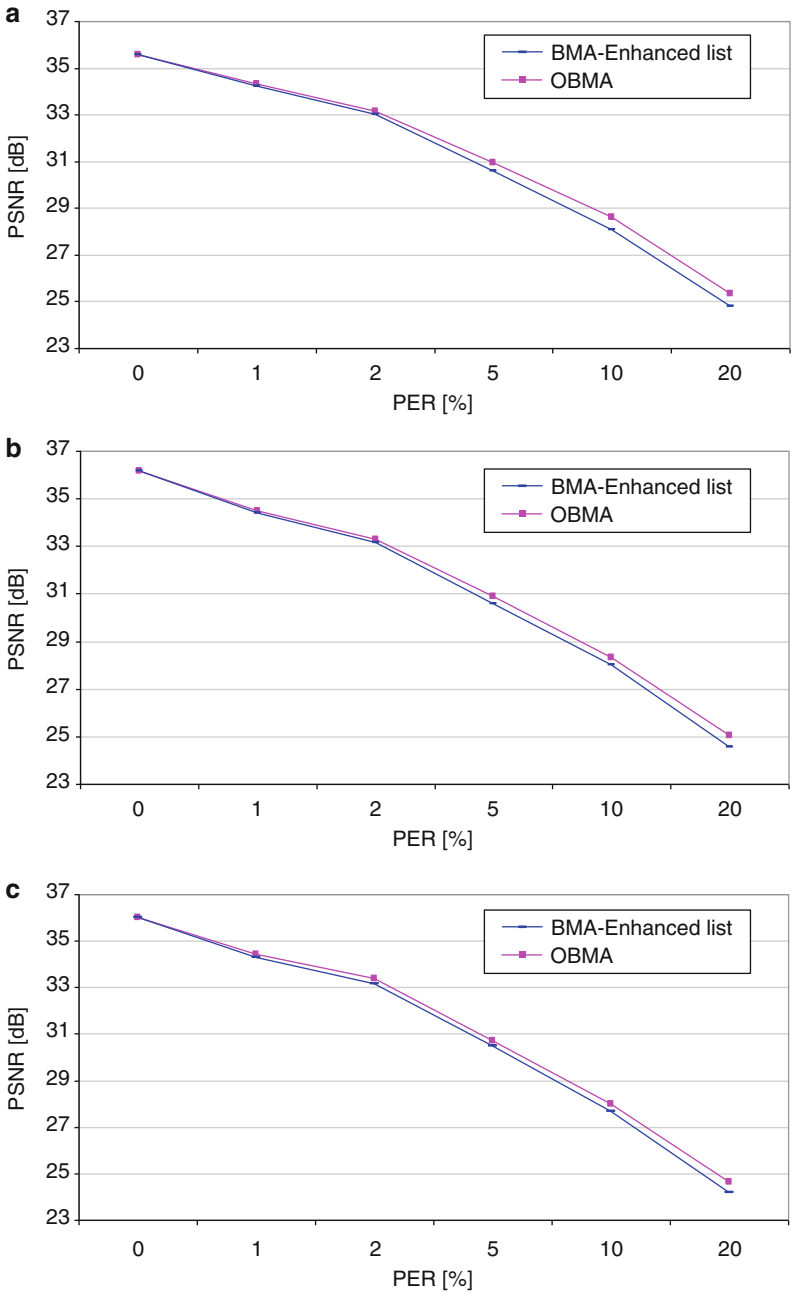


Fig. 12.20 Peak signal-to-noise ratio curves for the *Ballroom* sequence [14] using the enhanced list with the inner boundary matching algorithm and the outer boundary matching algorithm for (**a**) View 0, (**b**) View 1, and (**c**) View 2

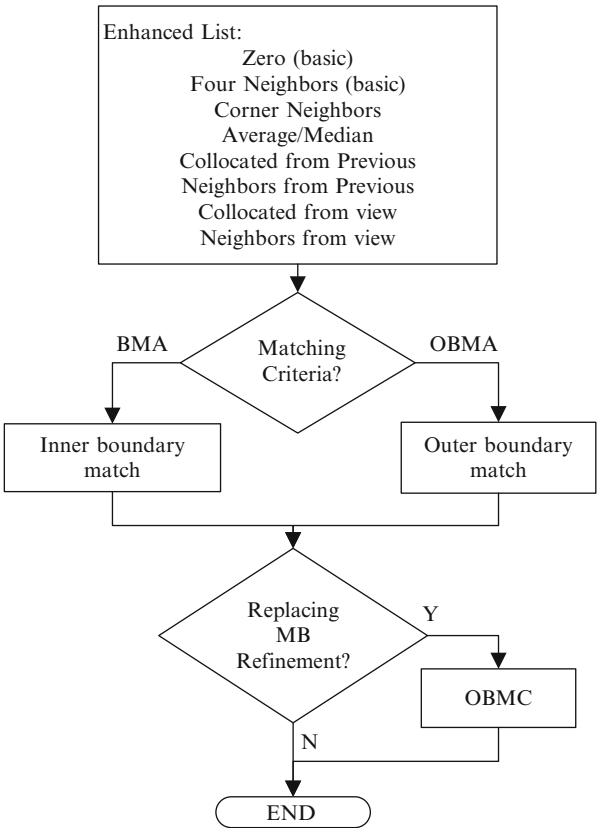


Fig. 12.21 The flowchart showing the steps involved in the advanced concealment technique. It starts with the evaluation of the enhanced list followed by the boundary matching selection criteria, and finally the application of the overlapped block motion compensation refinement

using the techniques discussed in the previous section. However, if the error affects an edge, geometric information is lost, leading to loss of the additional data that can be used for concealment. Furthermore, such errors will impinge on the reconstructed views. Therefore, techniques that better protect the edges in the depth videos need to be developed and advanced concealment methods in such regions are an important requirement.

12.6 Conclusion

In this chapter we have gone through a review of simple error resilient techniques and error concealment methods that can be applied to multi-view video coded sequences. Progressively more complex techniques were applied to present the reader with the improvements in obtained performance as more algorithms and

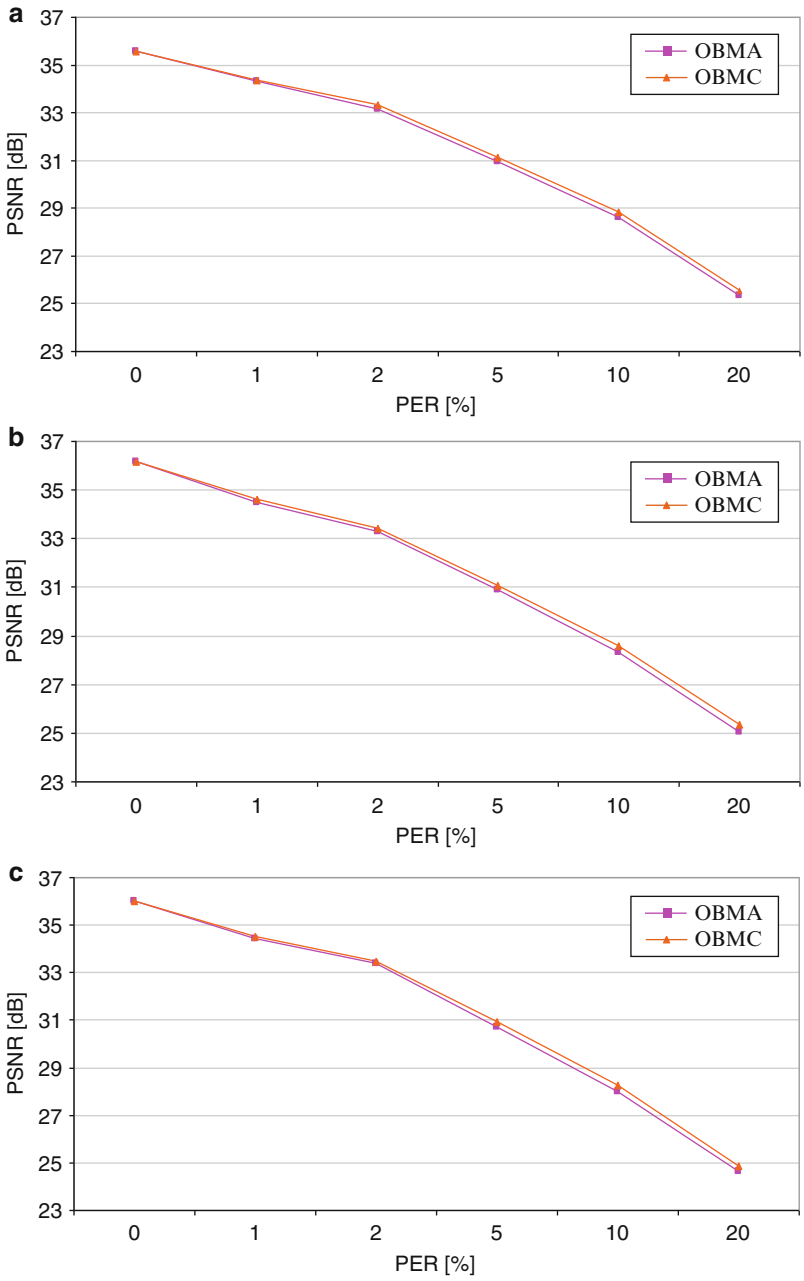


Fig. 12.22 Peak signal-to-noise ratio curves for the *Ballroom* sequence [14] using the enhanced list with the outer boundary matching algorithm with and without the overlapped block motion compensation algorithm for (**a**) View 0, (**b**) View 1, and (**c**) View 2

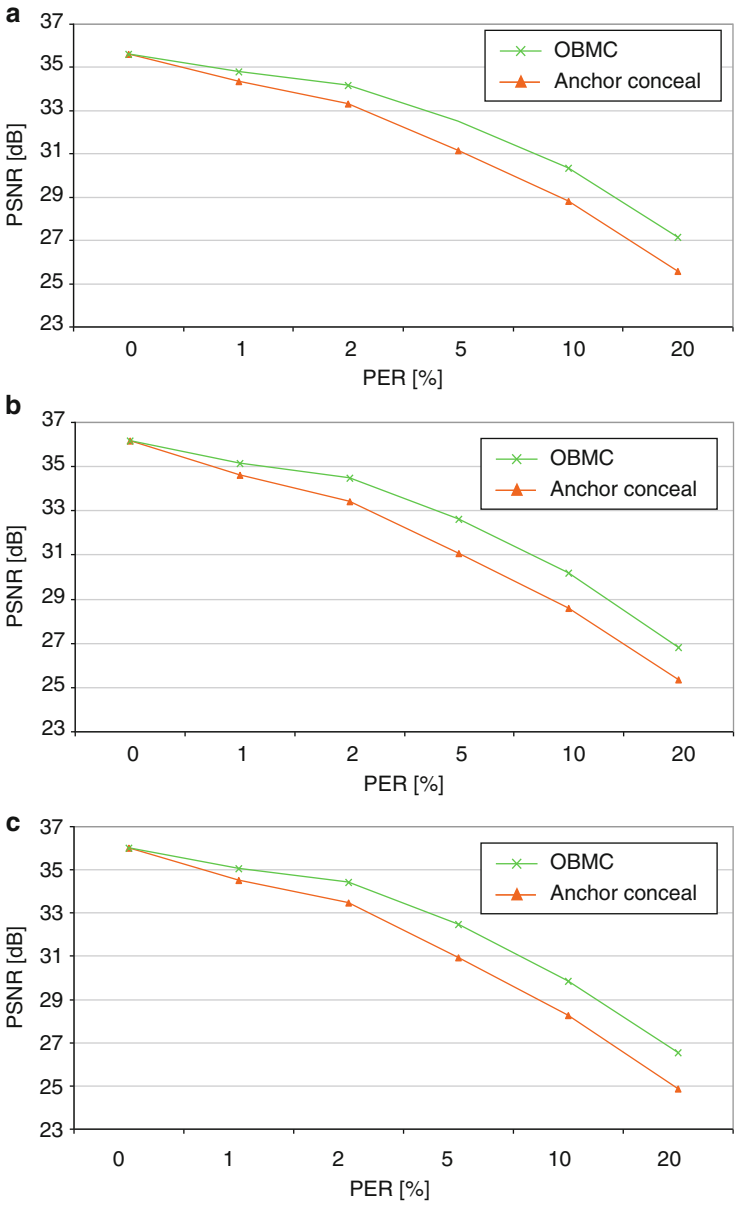


Fig. 12.23 Peak signal-to-noise ratio curves for the *Ballroom* sequence [14] using the anchor concealment method for (**a**) View 0, (**b**) View 1, and (**c**) View 2

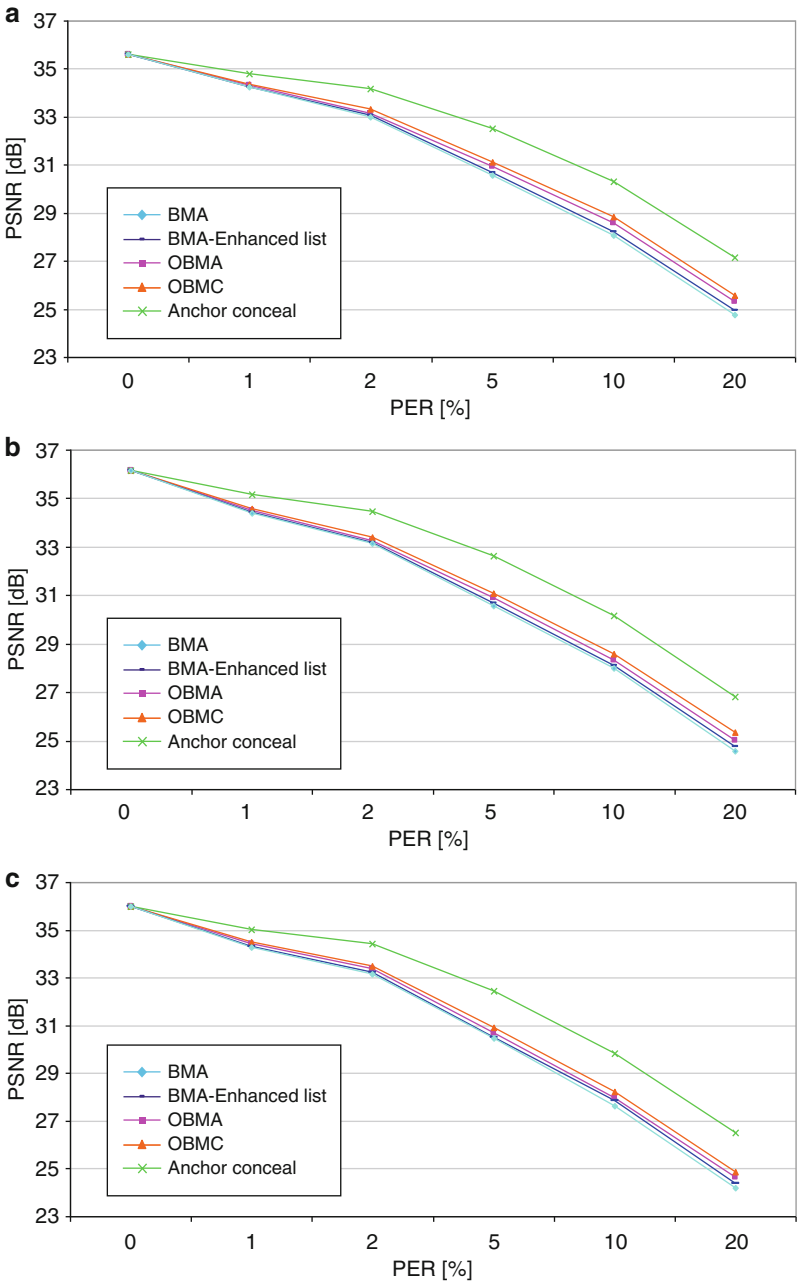


Fig. 12.24 Comparison of the achieved PSNR curves for (**a**) View 0, (**b**) View 1, and (**c**) View 2 of the *Ballroom* sequence and (**d**) View 0, (**e**) View 1, and (**f**) View 2 of the *Vassar* sequence [14] when using the different enhanced concealment methods

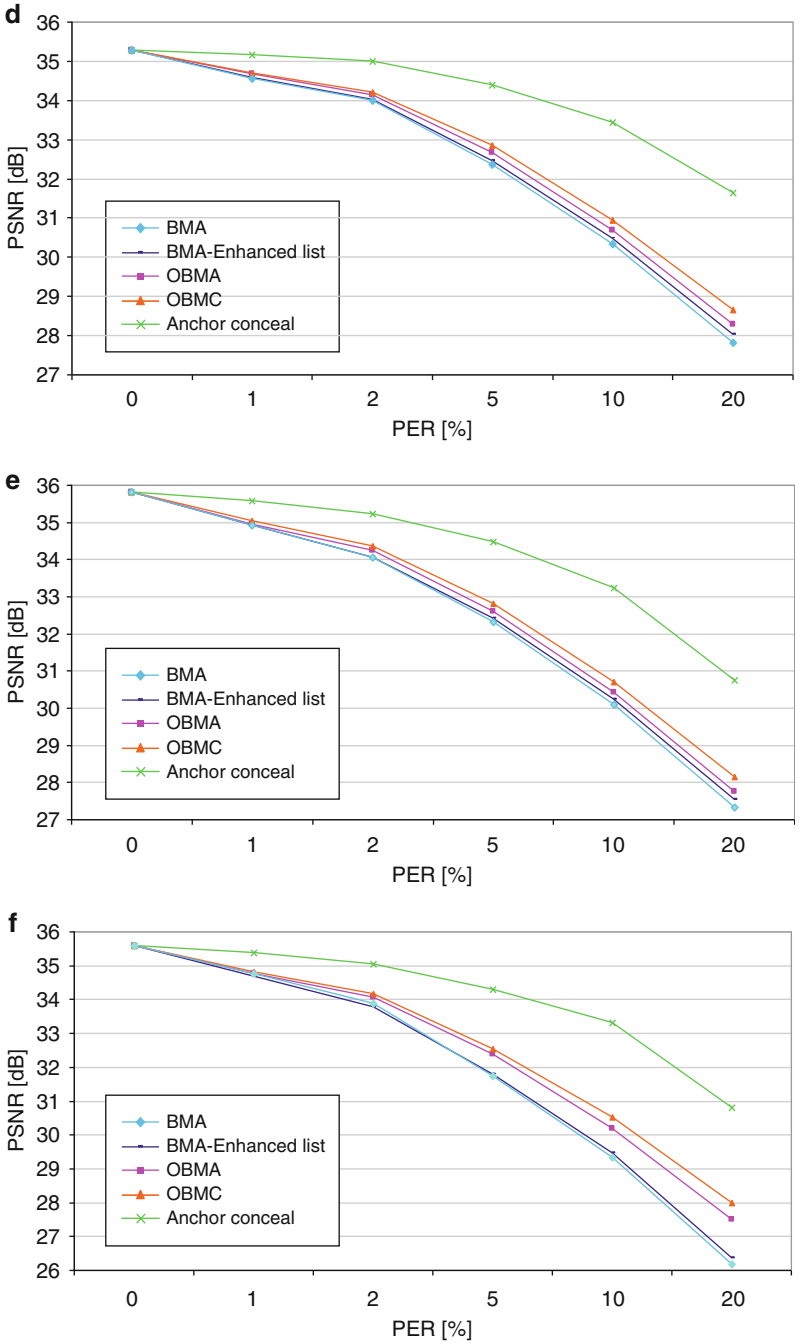


Fig. 12.24 (continued)

metrics are added to the concealment process. Further work still needs to be done in this field, coupled with better measurements of quality of experience for 3D video transmission, to ensure high-definition, good-quality 3D video display. This will ensure that more users will be keen to utilize the technology, allowing for the expansion of 3D multimedia services and applications.

Acknowledgments The authors would like to thank Dr Reuben Farrugia for the valuable discussions. They would also like to thank the Mitsubishi Electric Research Laboratory (MERL) [14] for providing a public repository with the test sequences for research efforts in this area.

References

1. Cisco Visual Networking Index (2012) Global mobile data traffic forecast update 2012–2017. In: White Paper, Cisco, February 2013
2. ITU-T Rec. H.264—ISO/IEC IS 14496–10 (2009) Advanced video coding for generic audio-visual services. March 2009
3. Mori Y, Fukushima N, Fujii T, Tanimoto M (2008) View generation with 3D warping using depth information for FTV. In: Proceedings of 3DTV conference, pp 229–232, May 2008
4. ISO/IEC MPEG, ITU-T VCEG (2007) Multi-view video plus depth (MVD) format for advanced 3D video systems. Document JVT-W100, April 2007
5. Kauff P, Atzpadin N, Fehn C, Müller M, Schreer O, Smolic A, Tanger R (2007) Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability. Signal Proc Image Commun 22(2):217–234
6. ISO/IEC MPEG, ITU-T VCEG (2011) Call for proposals on 3D video coding technology. Document N12036, March 2011
7. Merkle P, Smolic A, Müller K, Wiegand T (2007) Efficient prediction structures for multiview video coding. IEEE Trans Circuits Syst Video Technol 17(11):1461–1473
8. Micallef BW, Debono CJ (2010) An analysis on the effect of transmission errors in real-time H.264-MVC bit-streams. In: Proceedings of the 15th mediterranean electrotechnical conference, pp 1232–1236, April 2010
9. Ukai K, Howarth PA (2008) Visual fatigue caused by viewing stereoscopic motion images: background, theories, and observations. Displays 29(2):106–116
10. Vetro A, Tourapis A, Müller K, Chen T (2011) 3D-TV content storage and transmission. IEEE Trans Broadcast 57(2):384–394, Special Issue on 3D-TV Horizon: Contents, Systems and Visual Perception
11. Ozaktas HM, Onural L (eds) (2008) Three-dimensional television, capture, transmission, display. Springer, New York
12. Ho YS, Oh KJ (2007) Overview of multi-view video coding. In: Proceedings of the 14th international workshop in systems, signals and image processing, pp 5–12, June 2007
13. Richardson IEG (2003) H.264 and MPEG-4 video compression, video coding for next-generation multimedia. Wiley, West Sussex
14. MERL sequences. ftp://ftp.merl.com/pub/avetro/mvc-testseq/orig-yuv. Accessed 14 March 2013
15. Farrugia RA, Debono CJ (2008) A Robust error detection mechanism for H.264/AVC coded video sequences based on support vector machines. IEEE Trans Circuits Syst Video Technol 18(12):1766–1770
16. Van der Schaar M, Turaga DS, Stockhammer T (2006) Mpeg–4 beyond conventional video coding: object coding, resilience, and scalability. Syn Lect Image Video Multimed Proc 2(1):1–86

17. Dhondt Y, Lambert P (2004) Flexible macroblock ordering, an error resilience tool in H.264/AVC. In: Proceedings of the 5th FTW PhD symposium, faculty of engineering, Ghent University: paper no. 106, December 2004
18. Wang Y, Zhu Q (1998) Error control and concealment for video communications: a review. Proc IEEE 86(5):974–997
19. Wang Y-K, Hannuksela MM, Varsa V, Hourunranta A, Gabbouj M (2002) The error concealment feature in the H.26L test model. In: Proceedings of the international conference on image processing, pp 729–732, September 2002
20. Micallef BW, Debono CJ (2010) Error concealment techniques for multi-view video. In: Proceedings of the IFIP 2010 wireless days, October 2010
21. Lam WM, Reibman AR, Liu B (1993) Recovery of lost or erroneously received motion vectors. In: Proceedings of the international conference on acoustics, speech and signal processing, vol 5, pp 417–420, March 1993
22. Thaipanich T, Wu P-H, Kuo C-CJ (2008) Low-complexity video error concealment for mobile applications using OBMA. IEEE Trans Consum Electron 54(2):753–761
23. Micallef BW, Debono CJ, Farrugia RA (2011) Performance of enhanced error concealment techniques in multi-view video coding systems. In: Proceedings of the 18th international conference on systems, signals and image processing, pp 89–92, June 2011
24. Chen MJ, Chen LG, Weng RM (1997) Error concealment of lost motion vectors with overlapped motion compensation. IEEE Trans Circuits Syst Video Technol 7(3):560–563
25. Agrafiotis D, Bull DR, Canagarajah N (2006) Optimized temporal error concealment through performance evaluation of multiple concealment features. In: Proceedings of the international conference on consumer electronics, pp 211–212, January 2006

Part IV
3D Applications

Chapter 13
3D Robotic Surgery and Training at a Distance

Maria G. Martini, Chaminda T.E.R. Hewage, and Moustafa M. Nasralla

Abstract The usage of 3D images and video in medical surgery and training applications contributes in the provision of more natural viewing conditions and improved diagnosis and operation. This is enabled by the recent advances in 3D video capturing and display technologies, as well as advances in robotics and network technologies. The latest advances in robotic surgery enable the performance of many surgical procedures; in particular, recent 3D endoscopes have improved the performance of minimally invasive surgical procedures. Based on these advances, performing or visualizing in real-time surgical procedures at a distance can be envisaged. In this chapter, we present a review of 3D robotic surgery and tele-surgery applications and a performance evaluation of 3D robotic tele-surgery and training over wireless networks based on the long-term evolution (LTE) 3GPP standard. Different scheduling strategies are compared and results are analyzed in term of the resulting quality of experience for the surgeon.

13.1 Introduction

The latest advances in robotics, 3D video capture, and display technologies, together with the design and deployment of next-generation networks, will enable novel challenging applications, including 3D robotic surgery, also at a distance. Different medical areas would benefit from this, and in particular minimally invasive surgery is expected to lead in terms of potential benefits and applications.

Minimally invasive robotic surgery offers improved outcomes for patients, with shorter recovery times due to reduced pain and trauma. At the same time, it provides surgeons with a greater range of motion and access. Surgeons are now performing many procedures by using control probes that manipulate surgical

M.G. Martini (✉) • C.T.E.R. Hewage • M.M. Nasralla
Kingston University, London, UK
e-mail: m.martini@kingston.ac.uk

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_13, © Springer Science+Business Media New York 2014

instruments inserted through keyhole-size incisions, whereas incisions of several centimeters are required in conventional open surgery. Surgeons are guided in their maneuvers by images acquired through tiny camera probes.

Almost every surgical area is utilizing minimally invasive techniques, including urology, cardiology, thoracic, vascular, bariatric surgery, and neurosurgery.

The enhanced depth perception produced by recent 3D endoscopes has been demonstrated to improve the performance of minimally invasive surgical procedures. Three-dimensional imaging also facilitates the training of minimally invasive surgery and may lessen the learning curve of these technically demanding procedures [1]. The increased depth of field facilitates intricate minimally invasive surgical procedures [2–5]; it allows better recognition of tissue layers and may facilitate complex maneuvers such as laparoscopic suturing or knot tying [6]. Skill tests performed assessing laparoscopic suturing and knot tying demonstrated a 25 % increase in speed and accuracy of these laparoscopic tasks when utilizing a three-dimensional video system as compared to a standard two-dimensional endoscopic video system [7]. Three-dimensional video systems facilitate surgical tasks in general and benefits are particularly evident for inexperienced laparoscopic surgeons.

Several robotic surgical systems have been developed for minimally invasive surgery, including commercialized products such as Da Vinci [9, 10], by Intuitive Surgical, of Sunnyvale, Calif., with two or three arms equipped with surgical tools and an extra arm with a stereoscopic video camera probe, approved by the US Food and Drug Administration (FDA); ZEUS, by Computer Motion [11, 12], consisting of three interactive robotic arms, in addition to a voice-controlled endoscope and a console unit; and the MiroSurge system [13–15], a prototype developed by the German Aerospace Center (DLR). In such systems, visual data captured by the endoscopic camera is made available on different systems in parallel, including displays for the surgeon and operating room staff. For instance in [13] the stereo image stream is captured by a video server at a frame rate of 25 Hz and an image resolution of 768 $\times$ 576 (PAL) for the left and right images; the main display is directly fed by the video server via shared memory communication and further clients can be connected via Ethernet. A possible configuration integrates a polarization-based stereo display or an auto-stereoscopic display with eye tracking to process and display left and right images.

Remote tele-surgery appears as a natural extension, enabling surgical intervention and training at distance. Tele-surgery tests are reported in [9–11, 16, 17], where video was encoded according to the MPEG-2 standard, through DVTS uncompressed quality video, or through commercial video codecs from Polycom and Haivision (no mention in the articles if a standard compliant or proprietary codec was used). For 3D video, left and right views were transmitted separately and the trade-off delay quality was recognized as a major issue. State-of-the-art advancements on 3D video coding, including joint compression of left and right views or color and depth components, and 3D reconstructions were not considered or described in the tests performed. Only very recently some works on 3D reconstruction for this scenario have been presented [8, 20, 21]. Most of the current studies and experiments address transmission over high quality fixed

networks. These, however, are not always available. Tele-surgery is expected in fact to provide major benefits in particular in extreme fields such as battlefields, underwater, space, remote areas such as arctic regions, and disaster territories. In this case, the use of wireless technologies is a must due to the lack of wire line communication infrastructure. Tele-surgery over wireless environments poses unique challenges such as preserving reliability, meeting constraints on a maximum tolerable delay, and providing the required level of security. Investigated options include geosynchronous satellites, due to the provision of good data bandwidth, although drawbacks include non-ubiquitous coverage and in particular delay, impeding real-time surgery interaction. Other experiments focusing on 2D video tele-surgery proposed the exploitation of a wireless link enabled by an unmanned aircraft (UAV) [18, 19]. A small drone was sent to fly in circles above the operating area. Video from the camera near the robot was compressed into MPEG format and beamed to the plane, which relayed the feed to the monitor. At the same time, motion commands from the surgeon's console were bounced through the plane to the robot, which responded only a fraction of a second later, performing tasks such as tying suture knots. Due to the risk associated to the vulnerability of airborne communications, a solution considering a combination of connections, for instance, a fleet of UAVs, could be envisaged in this case.

Securing the network quality of service (QoS) is indeed crucial for critical real-time applications such as a remote surgery systems, For instance, in one of the described procedures, there was a significant amount of visual packet loss which obscured movement of the surgical arms, requiring the distant surgeon to act in a mentoring role, while the local surgeon completed the procedure.

In recent years, the concept of *QoS* has been extended to the new concept of *quality of experience* (QoE), as the first only focuses on the network performance (e.g., packet loss, delay, and jitter) without a direct link to the perceived quality, whereas the QoE reflects the overall experience of the user accessing and using the provided service. Experience is user- and context-dependent (involving considerations about subjective multimedia quality and users' expectation based on the cost they paid for the service, on their location, on the type of service, on the convenience of using the service, etc.).

It is then important that a remote surgery system is designed with the goal of meeting a required QoE determined by the users (medical specialists and trainees).

Lossless compression techniques are often considered in the medical imaging area, although these have the disadvantage of low compression rates. For instance, in [22] uncompressed transmission is considered for stereoscopic video signals acquired through the Da Vinci surgical system over a high bit rate connection. When transmission is over bandlimited and error-prone channels, a compromise must be made between compression fidelity and protection and resilience to channel errors and packet loss. It has been estimated that lossy compression ratios of 1:5 to 1:29 do not result in a lowering of diagnostic accuracy [25]. Furthermore, even in situations where the final diagnosis must be carried out on an image that has been reversibly compressed, irreversible compression can still play a critical role where quick access over a bandlimited channel is required [23, 24, 26, 27].

In this chapter, lossy video compression is considered for medical video transmission over wireless channels for telemedicine, surgery, and medical education purposes. Furthermore, simulcast/parallel encoding of left and right stereoscopic video is considered in this work. This enables independent control of left and right views. In order to provide better diagnostic quality, a low Quantization Parameter (QP) is employed. However, this configuration of 3D video encoding lacks the ability to exploit the redundancies present in the corresponding views. Possible stereoscopic video encoding architectures which can exploit inter-view redundancy are discussed in [28]. For instance, Scalable Video Coding (SVC) can be employed to encode left and right views in a backward compatible way. Even though these schemes provide some coding gain over simulcast configuration, the complexity of the encoder increases. Moreover, error propagation due to packet losses, not only along the sequence but also along the corresponding view, may result in visible impairments due to inter-frame dependencies. The requirement of good image quality for diagnosis and surgical procedures can hence be fulfilled by a simulcast configuration at low compression levels.

We consider in this chapter the case where 3D video sequences acquired through a robotic surgery system are transmitted real time over a wireless system, and we investigate the acceptability of the results in terms of QoE for the case of medical education and for wireless robotic tele-surgery.

The requirement of near real-time transmission derives from the need of interaction with the real scene in the case of surgery, and it is a desirable feature also when video is transmitted for education purposes (training of surgeons).

We refer to the latest 3GPP standard for wireless cellular communications, i.e., the long-term evolution (LTE) standard and its advanced version LTE Advanced (LTE-A). Specifically, we test the performance of real-time 3D surgery video transmission with different scheduling strategies and we compare the relevant results.

To the authors' best knowledge, this is the first work presenting a performance evaluation of 3D robotic surgery video transmission over a wireless system apart from preliminary results from the same team in [29]. The performance is not evaluated here only in terms of network performance, but the focus is in particular on the final received video quality and on its acceptability from the end user, i.e., the surgeon or trainee. Previous works on wireless transmission of medical video mainly focused on ultrasound video sequences (see, e.g., [23, 24, 26, 27, 30]).

The remainder of this chapter is organized as follows. Section 13.2 introduces the LTE wireless standard and presents the considered scenario with the coding (compression) and transmission schemes adopted. Section 13.3 describes the considered simulation setup, and Sect. 13.4 presents the results, both in terms of objective metrics and in terms of subjective quality as perceived by medical specialists.

13.2 LTE Wireless Systems and Proposed Transmission Strategy

13.2.1 The LTE Wireless Standard

In conjunction with the increasing growth of multimedia, internet, and real-time services, the LTE wireless standard has emerged to cope with these services efficiently. LTE has been introduced by the Third Generation Partnership Project (3GPP) as the next technology after 3.5G (HSPA+) cellular networks. The system architecture of the 3GPP LTE system contains several base stations called "eNodeB" (evolved node B) where the packet scheduling process is performed along with other radio resource management (RRM) tasks. LTE uses orthogonal frequency-division multiple access (OFDMA) in the downlink and single-carrier frequency-division multiple access (SC-FDMA) in the uplink. OFDMA extends the multicarrier technology OFDM to provide a better and more flexible multiple-access scheme. In other words, OFDMA splits the frequency band into multiple orthogonal subcarriers. This helps in improving the system capabilities by supporting high data rates, providing multiuser diversity, compacting the ISI (inter-symbol interference) factor, and creating immunity to frequency-selective fading of radio channels [31–33].

The QoS in the LTE downlink is affected by a significant number of factors such as channel conditions, available resources, and the type of application (e.g., delay sensitive/insensitive). In LTE the resources allocated to a user in the downlink are in the frequency and time domains. These resources are called physical resource blocks (PRB).

The channel bandwidth is divided into several 180 kHz PRBs. Each PRB consists of 6 or 7 OFDM symbols in the time domain and 12 consecutive orthogonal subcarriers in the frequency domain. Resource allocation is performed by every transmission time interval (TTI) whose duration is 1 ms.

Various packet scheduling algorithms have been developed to support real-time (RT) and non-real-time (NRT) flows, such as proportional fair (PF), modified largest weighted delay first (M-LWDF), and exponential PF (EXP/PF) [34–37]. In the aforementioned schedulers, each radio bearer is assigned a priority value by considering specific metrics. The bearer with the best metric is scheduled first at the next TTI.

13.2.2 Scheduling Strategies

Scheduling algorithms differ in the way they calculate the metric considered for the assignment of the resources to the different users.

In most of the cases, scheduling algorithms require the knowledge of the average transmission data rate ($\acute{R}_i$) of each flow associated to user i, and the instantaneous

available data rate for each sub-channel. As a result, when information about the performance guaranteed in the past for each flow is available, the system can perform fairness balancing. The following equation explains how $\acute{R}_i$ is typically estimated:

$$\acute{R}_i(f+1) = 0.8\acute{R}_i(f) + 0.2r_i(f) \tag{13.1}$$

$r_i(f)$ is the total achieved data rate by the i-th flow in the last scheduling epoch f (i.e., total number of bits transmitted over the entire PRBs allocated to i-th flow per TTI), $\acute{R}_i(f)$ is the average data rate achieved in the previous TTI, and $\acute{R}_i(f+1)$ is the average data rate achieved in the next TTI.

In what follows, three state-of-the-art downlink scheduling algorithms are described, with reference to the metrics considered and calculated.

The PF scheduler assigns radio resource blocks by taking into account both the experienced channel quality and the past user throughput [35]. The goal of this scheduler is to optimize the total network throughput and to guarantee fairness among the flows. Hence, the metric $\Lambda_{i,j}$ in this scheduler is defined as the ratio between the instantaneous data rate of the i-th flow in the j-th sub-channel (i.e., $r_{i,j}$) and the average data rate. The following illustrates the metric formula used in PF scheduler:

$$\Lambda_{i,j} = \frac{r_{i,j}}{\acute{R}_i} \tag{13.2}$$

where $r_{i,j}$ is computed based on the adaptive modulation and coding (AMC) scheme adopted, which is selected according to the channel quality indicator (CQI) feedback received from the UE. This feedback represents the channel quality (e.g., signal to interference plus noise ratio (SINR)) of the received i-th flow in the j-th sub-channel. $\acute{R}_i$ is the estimated average data rate.

The M-LWDF scheduler is developed to support multiple data users with different QoS requirements [36]. When real-time (RT) flows are considered, there must be a packet delay threshold t_i considered in its metric. Hence, this scheduler considers the best channel condition and the highest delay for the RT flows' head of line (HoL) in its allocation scheme. The following illustrates the metric used in the M-LWDF scheduler:

$$\Lambda_{i,j} = a_i D_{\mathrm{HoL},i} \frac{r_{i,j}}{\acute{R}_i} \tag{13.3}$$

where a_i is the maximum probability that the delay, D_{HoL}, of the head of line packet (i.e., the first packet to be transmitted in the queue) exceeds the target delay; therefore packets belonging to a RT service will be erased from the MAC queue if they violate the target delay. $r_{i,j}$ and $\acute{R}_i$ have the same meaning of the symbols in (13.2). For NRT services the metric used will be the one presented for the PF scheduler.

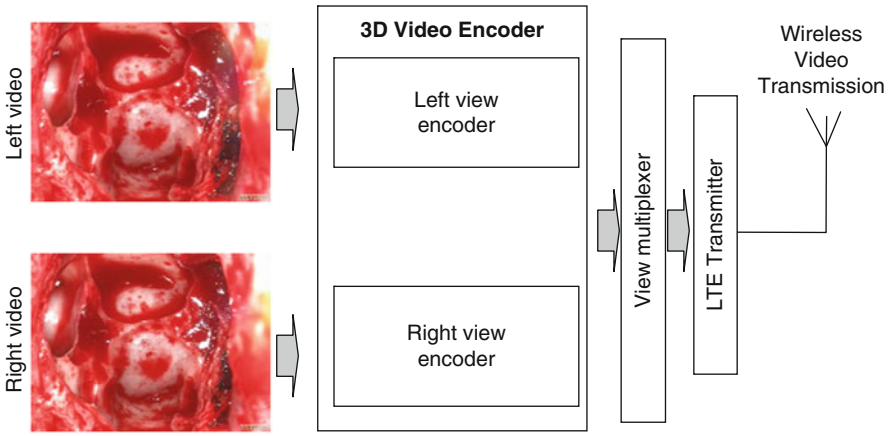


Fig. 13.1 Block diagram of the proposed 3D medical/surgical video transmission system

The EXP/PF scheduler is developed to maximize the priority of RT flows with respect to NRT ones. Also, this scheduler is designed to deal reliably with both RT and NRT users [37]. The following illustrates the metric formula used in EXP/PF scheduler:

$$\Lambda_{i,j} = \begin{cases} \exp\left(\dfrac{a_i D_{\mathrm{HoL},i-h}}{1+\sqrt{h}}\right)\dfrac{r_{i,j}}{\acute{\mathrm{R}}_i} \ldots\ldots i \in \mathrm{RT} \\ \dfrac{r_{i,j}}{\acute{\mathrm{R}}_i} \ldots\ldots i \in \mathrm{NRT} \end{cases} \tag{13.4}$$

$$h = \frac{1}{N_{rt}} \sum_{i=1}^{N_{rt}} a_i D_{\mathrm{HoL},i} \tag{13.5}$$

where symbols have the same meaning of the ones in (13.2) and (13.3) and N_{rt} is the number of active downlink RT flows.

13.2.3 Proposed Transmission Scheme

The proposed 3D medical/surgical video transmission system over LTE is illustrated in Fig. 13.1. Left and right image sequences captured from a 3D endoscope are fed into the 3D video encoder module for compression. The 3D video encoder consists of two video encoders running in parallel. The output of the 3D video encoder module is then fed into the multiplexer module. This module

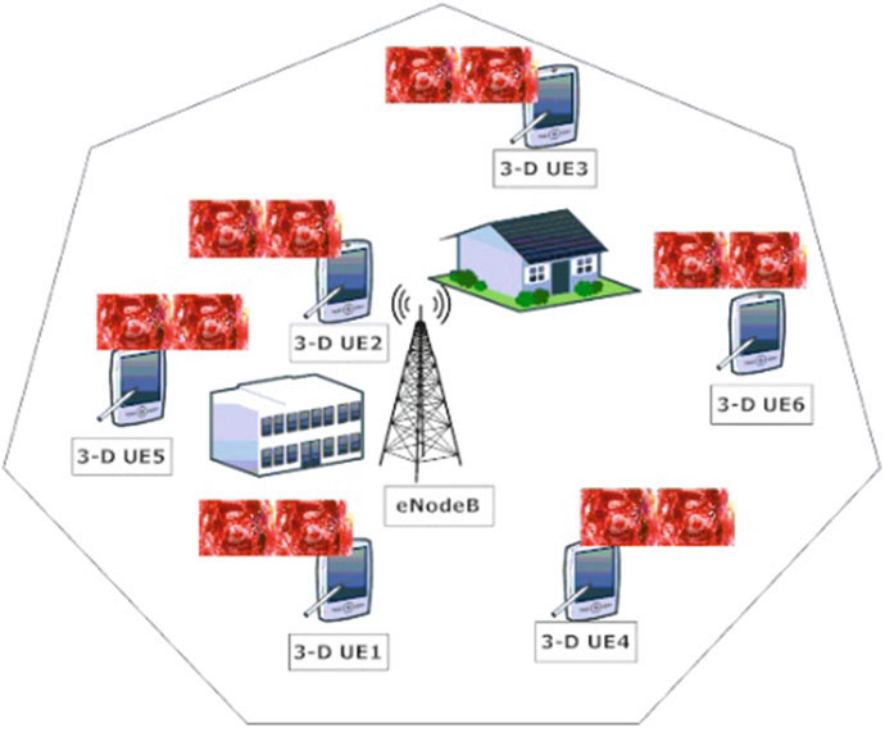


Fig. 13.2 3D video users over LTE radio network

multiplexes the encoded left and right bitstreams together and performs packetization. Finally the multiplexed packets are transmitted over the LTE network to the intended destination. Figure 13.2 depicts the network architecture used when 3D videos are transmitted over LTE to multiple users in the cell.

13.3 Simulation Setup

13.3.1 3D Video

Left- and right-based stereoscopic video content is considered for this investigation. The 3D medical video contents are provided by Visionsense Corp., USA. Visionsense produces a miniature stereoscopic (3D) sensor that optically maps the surgical field. Two sample left and right frames produced using the Visionsense 3D endoscope are shown in Fig. 13.3. These two selected sequences (acquired during different surgical procedures) are used in the experiments.

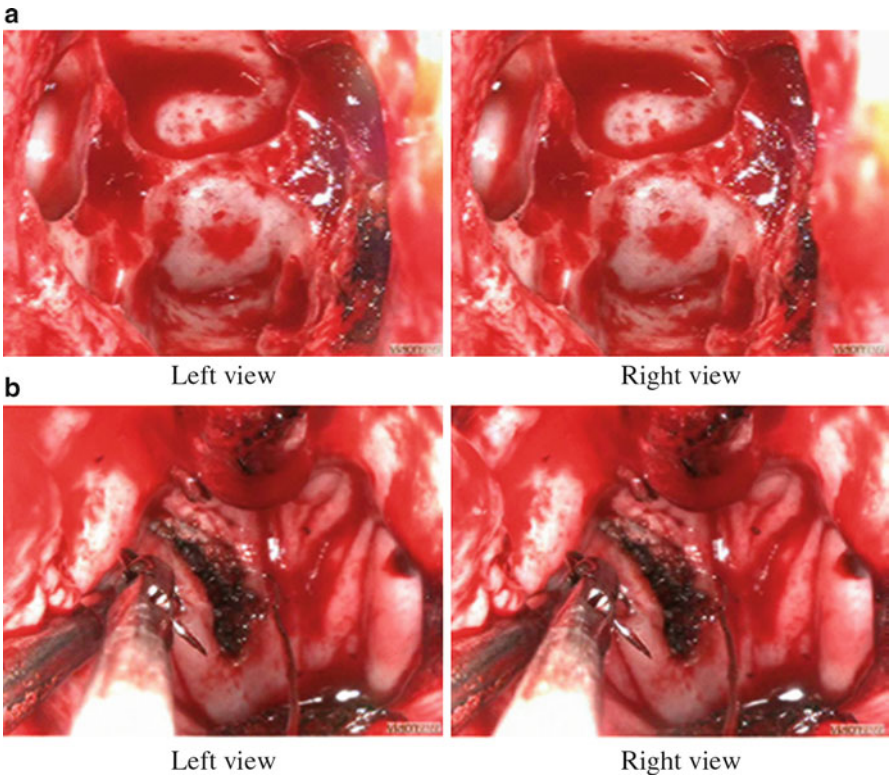


Fig. 13.3 Two sample 3D endoscope videos: (**a**) pituitary and (**b**) clivus biopsy suture

The 3D test sequences are encoded using the H.264/AVC video coding standard (JM reference software Version 16.0). Two codecs are employed to encode left and right views separately. Twenty-second long sequences (i.e., 1,000 frames at 50 fps) are encoded with IPPP. . .IPPP. . . sequence format, using QP value 30 for both I and P frames. An I frame is inserted by every 1 s of the video sequence to minimize the effect from temporal error propagation. Five hundred byte slices are also introduced in order to make the decoding process more robust to errors. Our previous study reported in [29] suggests that a 500 byte packet size is suitable to achieve high 3D video quality over LTE-like networks.

Our 3D video application of 2.2 Mbps is modeled based on a trace based strategy, i.e., we assume the application sends packets based on realistic video trace files obtained with the characteristics above.

In order to obtain stable results, simulations are run for several times (10) and the image quality is averaged over 1,000 frames. The corrupted bit-streams are later decoded using the JM reference software decoder. Slice copy is enabled during decoding to conceal the missing slices/packets of the corrupted bit-stream.

At each error condition, the left and right view quality is measured using the peak signal to noise ratio (PSNR) and SSIM quality metrics.

The received quality of 3D medical video content can be measured both objectively and subjectively. There is no reliable objective 3D video quality metric to date. However, studies show that individual objective quality measures of left and right views are highly correlated with perceptual 3D video quality scores (e.g., Mean Opinion Score (MOS)) [28, 38]. Therefore, in this work, the peak signal to noise ratio (PSNR) scores and SSIM scores of left and right views are considered as objective quality rating of the received surgical 3D video content.

13.3.2 LTE System

We investigate the performance of PF, M-LWDF, and EXP/PF when 3D traffic is transmitted over LTE systems.

The scenario used in this process is a single cell with interference. We have set up a scenario where there are 6 3D users. Users are randomly distributed with uniform distribution in the cell, with distance from the eNodeB ranging from 10 to 900 m. We assume users are moving at a pedestrian speed of 3 km/h in random directions (i.e., the speed direction is chosen randomly for each user, and it remains constant during the time and moves towards the boundary area. Once the user reaches the boundary area, the user chooses a new speed direction). The LTE propagation loss model is composed by four different models (path loss, penetration loss, shadowing, and multipath) [39].

The serving LTE base station (eNodeB) is located at the center of the cell. In the eNodeB a scheduler controls all the available PRBs by assigning them to the active flows which are competing for resources.

The LTE-Sim simulator is used to simulate this scenario [40]. In the time domain, resource allocation is performed by every TTI (i.e., every 1 ms). Two time slots of 0.5 ms compose one TTI which consists of 14 OFDM symbols with short cyclic prefix. Ten consecutive TTIs compose the LTE frame. Simulation parameters are shown in Table 13.1.

In the system, CQI feedback should be reported to the eNodeB by the users using the uplink control messages over the Physical Uplink Shared Channel (PUSCH), as this CQI value represents the users' instantaneous channel quality at each sub-channel. There are different feedback granularities in the standard; in our scenario the UE sends a single CQI for every resource block in the bandwidth. The mapping procedure among SINR and CQI is obtained using the mapping tables presented in [41]. Therefore, the selected MCS guarantees a robust communication and good service delivery. Moreover, this decision ensures that the estimated block error rate (BLER) remains under the target BLER (10 %).

Table 13.1 LTE downlink simulation parameters

Parameters	Values
Simulation duration	40 s
Flow duration	19.980 s
Frame structure	FDD
Cell radius	1 km
E-UTRA frequency band	1 (2.1 GHz)
Bandwidth	10 MHz
Number of PRBs	50
Slot duration	0.5 ms
Scheduling time	1 ms
Max delay	0.1 s
Average video bit rate	2201.19 kbps
Propagation loss/channel model	Typical urban/pedestrian A

13.4 Results

13.4.1 Network Performance

The network performance is assessed in terms of cell packet loss ratio and the cell throughput.

In order to compare results, the following abbreviations are used: "PF" represents the PF algorithm in (13.2), "M-LWDF" represents the M-LWDF algorithm in (13.3), and "EXP/PF" represents the EXP/PF algorithm in (13.4).

Figure 13.4 depicts the cell throughput when the cell is loaded increasingly by users starting from one 3D user up to six 3D users. The figure shows that the M-LWDF scheduler outperforms the EXP/PF and PF schedulers. This means the M-LWDF scheduler has the capability to cope with RT services better than the others and it shows good performance when the number of UEs increases.

Figure 13.5 depicts the cell packet loss ratio. Again the M-LWDF scheduler produces better PLR values when different scenarios are used. As the number of UEs increases, the load on the serving LTE cell increases and the packet loss ratio increases. We can observe from Fig. 13.5 that the packet loss ratio is not acceptable when we have more than four 3D users in the cell. The corresponding value is 10 %. This value is mostly considered as acceptable for video traffic by various wireless operators. The quality level varies from one scheduler to another and the packet loss ratio of the M-LWDF scheduler is lower than the others. Therefore, the M-LWDF scheduler is the most recommended in this scenario.

Figure 13.6 represents the cell spectral efficiency [bits/s/Hz]. The higher the spectral efficiency, the higher the channel utilization and the higher the throughput. Hence, as noticed from the figure, the M-LWDF scheduler utilizes the channel more fficiently than the EXP/PF and PF schedulers. Therefore, the throughput, the spectral efficiency, and the system performance achieved by the M-LWDF are higher than those for the other two schedulers.

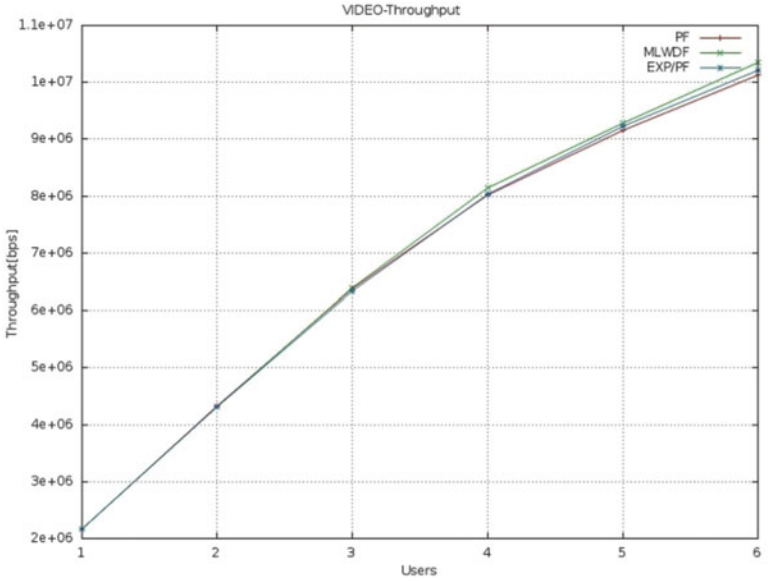


Fig. 13.4 3D video cell throughput

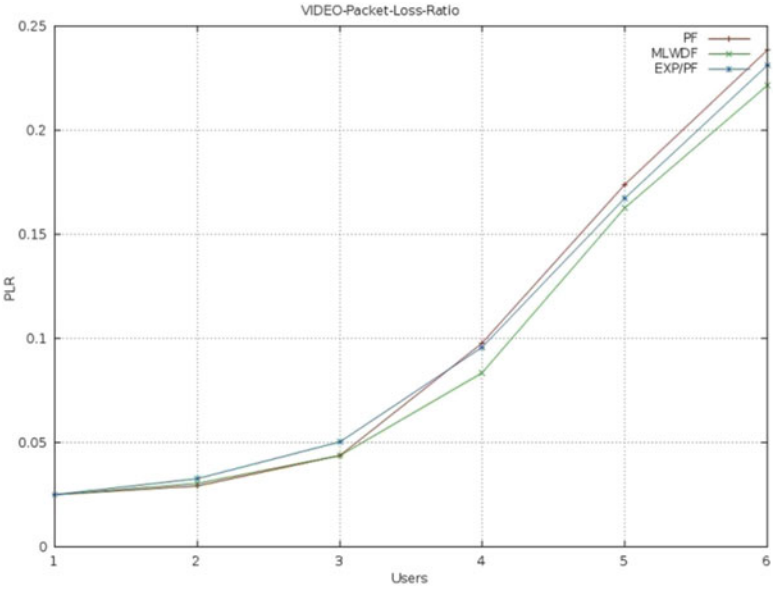


Fig. 13.5 3D video cell packet loss ratio

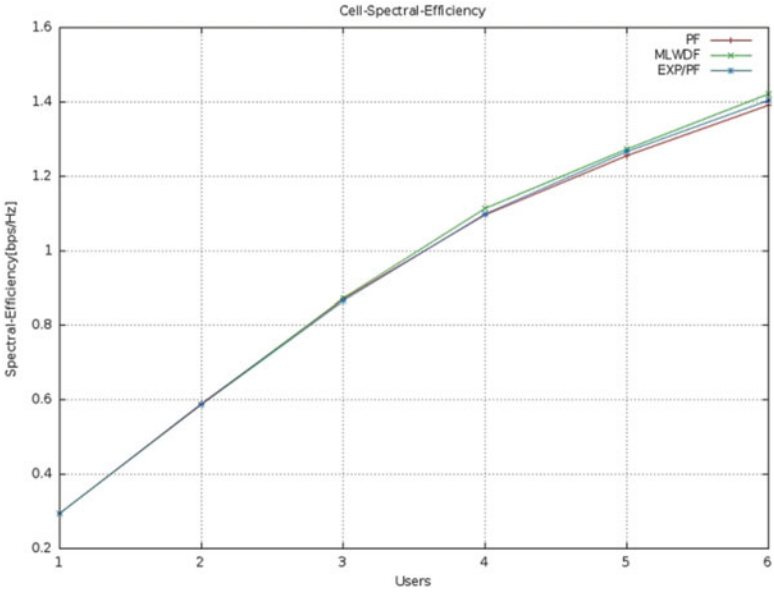


Fig. 13.6 Cell spectral efficiency

13.4.2 3D Video Quality Assessment

The performance of 3D tele-robotic surgery and training using three different scheduling algorithms for LTE is evaluated using both objective and subjective quality evaluation methods for the received 3D video. Quality evaluation studies are carried out using the sequence captured during the "Pituitary" surgical procedure. The average PSNR and SSIM of left and right video are considered as a measure of objective quality of 3D video. The correlation between average left and right PSNR/SSIM vs. true 3D perceptual quality is discussed in [38]. The average PSNR and SSIM image quality results for the scenarios of "4 Users" and "5 Users" are shown in Tables 13.2 and 13.3, respectively. According to Tables 13.2 and 13.3, most of the users achieve reasonable PSNR and SSIM image quality ratings. The average received image quality is better in the scenario of "4 Users" compared to "5 Users." User 2, 4, and 5 in Table 13.3 achieve a lower quality left and right image (less than 30 dB), and this may not be suitable for clinical applications. The reduced quality in the "5 Users" scenario is mainly due to the scarce resources in the LTE transmission system for the given bandwidth (10 MHz).

Figures 13.7 and 13.8 demonstrate the average image quality for the "4 Users" case as measured by PSNR and SSIM, respectively. It can be observed that the "M-LWDF" scheduling algorithm achieves better image quality compared to "EXP/PF" and "PF" methods. Therefore, it can be concluded that the "M-LWDF" scheduling method provides improved 3D image quality over next-generation wireless networks such as LTE.

Table 13.2 Average PSNR and SSIM quality results for "4 Users"

"4 Users"	Average left and right PSNR (dB)			
Method	User 1	User 2	User 3	User 4
M-LWDF	31.905	34.075	31.905	36.520
EXP/PF	31.010	32.820	31.175	36.520
PF	29.640	32.950	30.830	36.520
"4 Users"	Average left and right SSIM			
Method	User 1	User 2	User 3	User 4
M-LWDF	0.8985	0.9194	0.9026	0.9383
EXP/PF	0.8824	0.9149	0.8869	0.9383
PF	0.8715	0.9159	0.8834	0.9383

Table 13.3 Average PSNR and SSIM quality results for "5 Users"

"5 Users"	Average left and right PSNR (dB)				
Method	User 1	User 2	User 3	User 4	User 5
M-LWDF	33.82	28.07	31.77	27.57	27.18
EXP/PF	33.98	26.28	30.74	27.88	27.64
PF	33.38	24.83	31.88	27.31	27.66
"5 Users"	Average left and right SSIM				
Method	User 1	User 2	User 3	User 4	User 5
M-LWDF	0.92	0.84	0.91	0.85	0.85
EXP/PF	0.92	0.83	0.90	0.86	0.85
PF	0.91	0.81	0.90	0.85	0.85

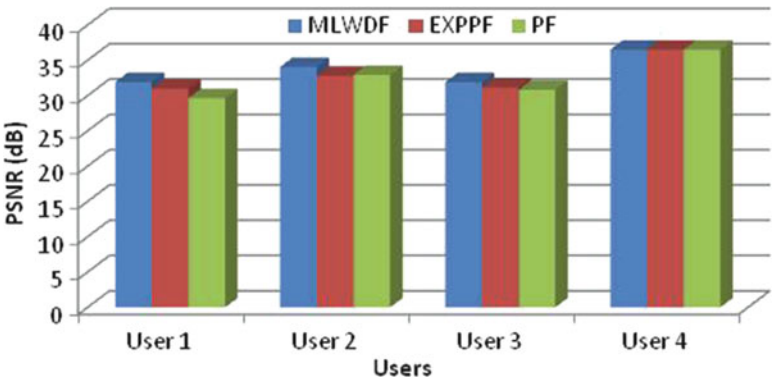


Fig. 13.7 Average PSNR (dB) image quality for 4 Users

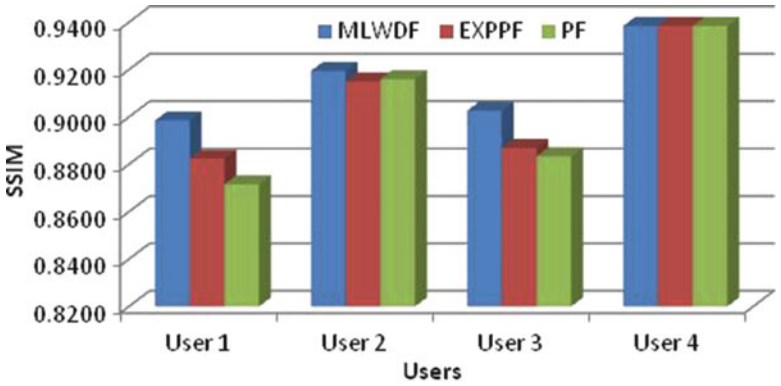


Fig. 13.8 Average SSIM image quality for 4 Users

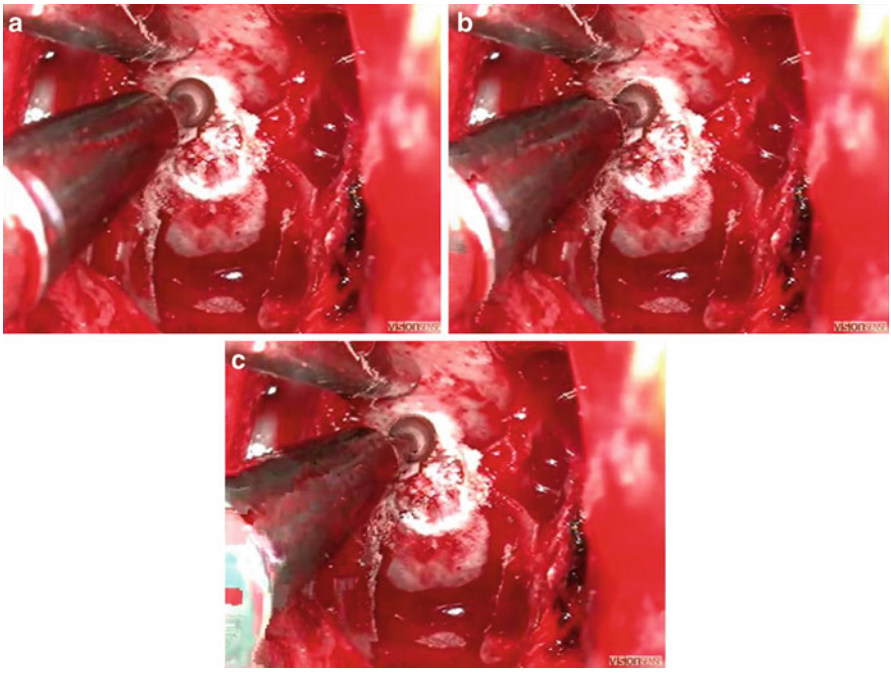


Fig. 13.9 Sample left image frames of User 2, "4 Users" scenario with (**a**) "M-LWDF" method, (**b**) "EXP/PF" method, and (**c**) "PF" method

Figure 13.9 reports sample left image frames received, with different scheduling algorithms, by User 2 in the scenario with "4 Users" in the system. According to these subjective illustrations, the superiority of the "M-LWDF" method over other conventional scheduling algorithms is evident.

13.5 Conclusion

The chapter presented a performance evaluation of 3D robotic surgery video transmission over a wireless network. Three different scheduling strategies have been compared. The performance was evaluated in terms of network parameters and objective video quality metrics. The work presented also enables a mapping among the different quality metrics used, from network parameters to image quality metrics.

It has been shown that the M-LWDF scheduling strategy results in the best performance with all the evaluation criteria and, for the case of an LTE system with 10 MHz bandwidth, a maximum of four users can be supported with acceptable quality. The results also suggest that it is expected that increasing the fairness of the scheduling algorithm further would result in a globally improved quality for the users in the cell.

References

1. Durrani AF, Preminger GM (1995) Three-dimensional video imaging for endoscopic surgery. Comput Biol Med 25(2):231–247
2. Satava RM (1993) 3-D vision technology applied to advanced minimally invasive surgery system. Surg Endosc 7:429–443
3. Birkett DH (1993) 3-D imaging m gastrointestinal laparoscopy. Surg Endosc 7:556–557
4. Jourdan I, Dutson E, Garcia A et al (2004) Stereoscopic vision provides a significant advantage for precision robotic laparoscopy. Br J Surg 91(7):879–885
5. Smith R, Day A, Rockall T et al (2012) Advanced stereoscopic projection technology significantly improves performance of minimally invasive surgical skills. Surg Endosc 26:1522–1527
6. Janetschek G, Reissial A, Peschel H et al (1993) Chip on a stick technology: first clinical experience with this new video-laparoscope. J Endourol 7:S195
7. Bahayan RK, Chiu AW, Este-McDonald J, Birkett DH (1993) The comparison between 2-dimensional and 3-dimensional laparoscopic video systems in a pelvic trainer. J Endourol 7:S195
8. Mountney P, Stoyanov D, Yang G-Z (2010) Three-dimensional tissue deformation recovery and tracking. IEEE Signal Process 27(4):14–24
9. Sterbis JR, Hanly EJ, Barry C et al (2008) Transcontinental telesurgical nephrectomy using the da Vinci robot in a porcine model. Urology 71:971–973
10. Marescaux J, Leroy J, Gagner M et al (2001) Transatlantic robot-assisted telesurgery. Nature 413(6854):379–380
11. Ballantyne GH (2002) Robotic surgery, telerobotic surgery, telepresence and telemonitoring. Surg Endosc 16:1389–1402
12. Holt D, Zaidi A, Abramson J, Somogyi R (2004) Telesurgery: advances and trends. Univ Toronto Med J 82(1):52–54
13. Hagn U, Konietschke R, Tobergte A et al (2010) DLR MiroSurge: a versatile system for research in endoscopic telesurgery. Int J CARS 5:183–193
14. Konietschke R, Tobergte A, Preusche C et al (2010) A multimodal training platform for minimally invasive robotic surgery. In: Proc. of 19th IEEE international symposium on robot and human interactive communication, Viareggio, Italy, Sept 12–15 2010

15. Bauernschmit R et al (2009) On the role of multimodal communication in telesurgery systems. In: Proc. of IEEE MMSP 2009
16. Arata J, Takahashi H, Pitakwatchara P et al (2006) A remote surgery experiment between Japan-Korea using the minimally invasive surgical system. In: Proc. of the 2006 I.E. international conference on robotics and automation, Orlando, Florida, May 2006
17. Arata J, Takahashi H, Pitakwatchara P et al (2007) A remote surgery experiment between Japan and Thailand over Internet using a low latency CODEC system. In: Proc. of the 2007 I.E. international conference on robotics and automation, Rome, Italy, Apr 2007
18. Rosen J, Hannaford B (2006) DOC at a distance. In: IEEE spectrum, Oct 2006
19. Lum M, Friedman D, King HH et al (2008) Teleoperation of a surgical robot via airborne wireless radio and transatlantic internet links. Field Serv Robot 42:305–314
20. Bouma H, van der Mark W, Eendebak PT et al (2012) Streaming video-based 3D reconstruction method compatible with existing monoscopic and stereoscopic endoscopy systems. Proc SPIE 8371:12
21. Zhou J, Zhang Q, Li B, Das A (2010) Synthesis of stereoscopic views from monocular endoscopic videos. In: Proc. 2010 I.E. computer society conference on computer vision and pattern recognition workshops (CVPRW), pp 55–62
22. Navratil J, Sarek M, Ubik S et al (2011) Real-time stereoscopic streaming of robotic surgeries. In: IEEE 13th Int. Conf. on e-Health networking, applications, and services
23. Martini MG, Istepanian RSH, Mazzotti M, Philip N (2010) Robust multi-layer control for enhanced wireless tele-medical video streaming. IEEE Trans Mob Comput 9(1):5–16
24. Istepanian RSH, Philip N, Martini MG (2009) Medical QoS provision based on reinforcement learning in ultrasound streaming over 3.5G wireless systems. IEEE J Sel Areas Commun 27 (4):566–574
25. Cosman PC, Davidson HC, Bergin CJ et al (1994) Thoracic CT images: effect of lossy image compression on diagnostic accuracy. Radiology 190:514–524
26. Martini MG, Papadopoulos H (eds) (2009) Health and inclusion. In: Strategic applications agenda. eMobility European Technology Platform
27. Martini MG (2008) Wireless broadband multimedia health services: current status and emerging concepts. IEEE personal indoor and mobile radio communications (PIMRC 2008), Cannes, France, Sept 2008
28. Hewage CTER, Karim HA, Worrall S et al. (2007) Comparison of stereo video coding support in MPEG-4 MAC, H.264/AVC and H.264/SVC. In: Proceedings of visual information engineering-VIE07, London
29. Hewage CTER, Martini MG, Khan N (2011) 3D medical video transmission over 4G networks. In: 4th international symposium on applied sciences in biomedical and communication technologies, Barcelona, Spain, Oct 26–29 2011
30. Istepanian RSH, Philip N, Martini MG et al (2008) Subjective and objective quality assessment in wireless teleultrasonography imaging. In: IEEE international conference engineering in medicine and biology society (EMBC08), Vancouver, August
31. Sesia S, Toufic I, Baker M (2009) LTE—the UMTS long term evolution. Wiley, UK
32. Tong W, Sich E, Peiying Z, Costa JM (2008) True broadband multimedia experience. IEEE Microwave Mag 9(4):64–73
33. Shakkottai S, Rappaport TS, Karlsson PC (2003) Cross-layer design for wireless networks. IEEE Commun Mag 41(10):74–80
34. Ramli H, Basukala R, Sandrasegaran K, Patachaianand R (2009) Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system. In: Proc. IEEE 9th Malaysia international conference on communications, pp 815–820
35. Choi JG, Bahk S (2007) Cell-throughput analysis of the proportional fair scheduler in the single-cell environment. IEEE Trans Veh Technol 56(2):766–778
36. Ameigeiras P, Wigard J, Mogensen P (2004) Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA. IEEE Trans Veh Technol Conf 2:999–1003

37. Basukala R, Mohd Ramli H, Sandrasegaran K (2009) Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system. In: Proc. IEEE F. Asian Himalayas Conf., Nov 2009
38. Hewage CTER, Worrall ST, Dogan S et al (2009) Quality evaluation of color plus depth map-based stereoscopic video. IEEE J Sel Top Signal Process 3(2):304–318
39. Tech. Specif. Group Radio Access Network 3GPP (2006). Physical layer aspect for evolved universal terrestrial radio access (UTRA) (Release 7). Technical report, 3GPP TR 25.814
40. Piro G, Grieco LA, Boggia G et al (2011) Simulating LTE cellular systems: an open source framework. IEEE Trans Veh Technol vol.60, no.2, pp.498,513, Feb. 2011
41. 3GPP Technical Specification 36.213. Physical layer procedures (Release 8). http://www.3gpp.org

Chapter 14
Future of 3DTV Broadcasting: The MUSCADE Perspective

Hemantha Kodikara Arachchi

Abstract Even though 3D television (3DTV) services are practically up and running and the compatible displays are readily available in the market, the technology is far from mature enough for gaining widespread acceptance from the users. In light of this experience, the MUSCADE project was set up to define, develop, validate, and evaluate the technological innovations in 3DTV capturing, data representation, compression, transmission, and rendering required for a technically efficient and commercially successful 3DTV broadcast system. This chapter summarizes the technologies developed within the MUSCADE consortium to shape the future 3DTV broadcasting.

14.1 Introduction

History of 3D imaging can be traced back to the beginning of photography. Scottish inventor and writer David Brewster has been credited for inventing the stereoscope that could take pictures in 3D in 1844. Louis Jules Duboscq succeeded in displaying a 3D picture of Queen Victoria at The Great Exhibition in 1851. He used an improved version of the stereoscope to produce the exhibit. In 1855, the Kinematoscope, a stereo animation camera, was invented. In the late 1890s British film pioneer William Friese-Greene filed a patent for a 3D movie process, marking the beginning of the stereoscopic motion picture era. However, his invention was far from practical because the movie has to be watched through a mechanism that converge side-by-side projected stereo images. This obstructive mechanism was not welcomed in movie theatres. In 1915, Astor Theatre in New York City made the

H. Kodikara Arachchi (✉)
I-Lab Multimedia Communications Research, University of Surrey, Guildford GU2 7XH, UK
e-mail: h.kodikaraarachchi@surrey.ac.uk

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1_14, © Springer Science+Business Media New York 2014

history for screening 3D video when Edwin S. Porter and William E. Waddell showed a set of test movie clips recorded in red-green anaglyph format. A paying audience got the chance of watching a 3D film for the first time in 1922 when anaglyph 3D movie "The Power of Love" was presented at the Ambassador Hotel Theater in Los Angeles. Later in the same year, Laurens Hammond and William F. Cassidy unveiled the earliest alternate-frame sequencing form of film projection. Known as the Teleview system, which projected left and right frames alternatively one after another in rapid succession, it relied on a special viewing device attached to the armrest of the seat for correctly presenting each view to viewer's eye.

While significant activities and breakthroughs were happening in 3D cinema frontier, a system for transmitting visual images over a greater distance and reproducing them was also being developed. The invention of television is credited to John Logie Baird, who demonstrated the technology to televise moving images in 1926. Two years after this landmark demonstration, he also managed to demonstrate the stereoscopic 3D television technology for the first time on his company's premises in London in 1928. He went on experimenting with a number of 3DTV systems using electromechanical and cathode-ray tube techniques.

WRGB is credited to be the world's oldest television station, which started its experimental broadcasting from the General Electric factory in Schenectady, NY, in 1928. However, no attempt has been reported to broadcast any 3D programs until recent time when the 3D television sets finally managed to creep up into the home audience. This is mainly fuelled by the availability of affordable 3D-enabled television sets across the consumer market. In line with this market trend and with the aim of emulating the unparalleled box-office success of a number of 3D movies [1], a number of television service providers inaugurated dedicated channels for 3D enthusiasts. Before that some leading television broadcaster began to offer limited number of 3D programs over existing 2D television channels. The first broadcaster who stepped into the 3D television era is the Japanese cable channel BS 11, which started providing 3D television programs in 2008. The target of BS 11 was to provide 3D programs at least four times a day. Many other content providers and broadcasters such as ESPN, DirecTV, Orange, and BSkyB launched 3D programs in 2010. 2010 became another landmark year for 3D television history when South Korea launched the first dedicated 3D channel, Sky 3D. This channel is broadcasted in side-by-side format in 1920x1080i resolution. Following this lead, a number of more dedicated 3D channels have been launched across the globe.

Even though 3DTV services are practically up and running and compatible displays are readily available in the market, the technology is far from mature enough for gaining widespread acceptance from the home audience. Essentially, the user experience has a profound impact on the attractiveness and sustainability of 3D television services. The present generation of displays and broadcasting technologies, which only provide stereoscopic video, have failed to convince the audience to take up 3D television due to poor experience. Hence, a technological revamp is required to improve the user experience in order to gain a sustainable audience for 3DTV channels.

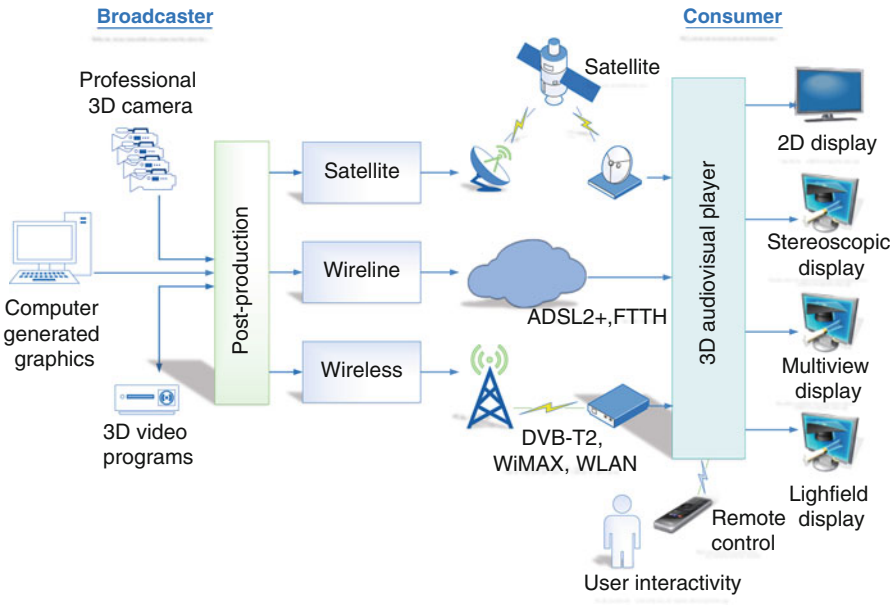


Fig. 14.1 MUSCADE 3DTV service architecture

In light of this experience, a consortium of twelve leading European industrial, research, academic and public institutions got together to set up the MUSCADE project [2] for raising the 3D television experience for the future generations. It is a large-scale integrated project funded by the European Union Framework Program 7. The mission of the MUSCADE project is to define, develop, validate, and evaluate the technological innovations in 3DTV capturing, data representation, compression, transmission, and rendering required for a technically efficient and commercially successful 3DTV broadcast system. The overall system architecture of the MUSCADE 3DTV service is shown in Fig. 14.1. The focus of this chapter is to briefly introduce major technological aspects investigated under the MUSCADE project for taking 3D broadcasting one step ahead of present frame-compatible stereoscopic infrastructures.

14.2 The State-of-the-Art 3DTV

At present, HDTV programs are broadcasted using terrestrial, satellite, cable, and broadband infrastructures. The terrestrial DVB-T and DVB-T2 offer the lowest bandwidth. DVB-T offers bandwidth in the range of 20–24 Mbit/s depending on

whether they are operated in single-frequency network (SFN) or multifrequency network (MFN) topology. DVB-T2 can extend the capacity up to 40 Mbit/s range. In general, up to 2 HDTV services can be accommodated in a single DVB-T multiplex while DVB-T2 has sufficient capacity for up to five channels. Moreover, typical DVB-S2 satellite networks offer a capacity of over 45 Mbit/s over a single 36 MHz transponder. Therefore each of such transponder offers enough capacity for more than five good-quality HD channels. Aggregating the capacities of all the transponders, a conventional satellite system can support a large number of high-quality HD channels.

Meanwhile, broadband infrastructure is fast becoming the favorite medium for delivery since modern access networks, which are often the bottleneck, now offer enough and consistent capacity to stream HD channels to home audience. A typical IPTV service needs maximum of only a 10 Mbit/s bandwidth budget for successful operation. It should be noted that this capacity has been estimated based on one high-quality HD service per household. A number of access technologies that have been used by present broadband service providers comfortably exceed this limit. Most of the present access networks are based on ADSL2+, which offers 24 Mbit/s theoretical downlink capacity. However, the optical infrastructures, which are being laid down across the globe, offer far better capacities. Hence, future of the IPTV delivery is more promising.

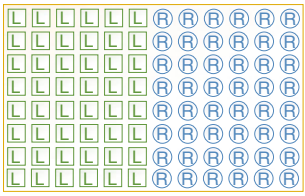
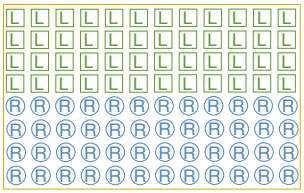
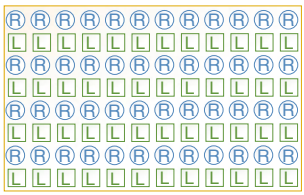
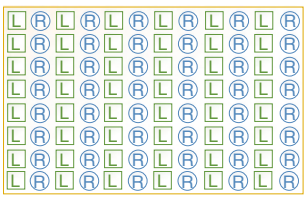
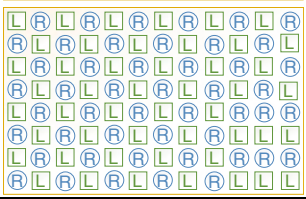
Present IPTV delivery infrastructures have been built around managed network concept to ensure efficient delivery of multicast video traffic. The present tradition is to multicast live television programs as an MPEG-2 Transport Stream. However, Video on Demand services, which are becoming popular, is provided as IP unicasts.

Present generation of 3DTV channels share the same HDTV infrastructures. Therefore, the broadcasters have no more bit budget for delivering 3D programs than that is allocated for HDTV broadcasting. Furthermore, in some cases, the home set-top box designed for HDTV reception has to be utilized for 3DTV reception since replacing set-top boxes is not feasible due to the cost. Considering these constraints, HD frame-compatible format has been adapted for 3DTV broadcasting. Since present HD compatible receivers can receive frame-compatible format, this format is suitable until 3D services are ubiquitous for justifying any investments for better infrastructure. Table 14.1 summarizes common frame-compatible formats for 3DTV broadcasting [3].

The disadvantage of the present frame-compatible stereoscopic format is that it offers very limited immersive experience. The spatial resolution of each view is reduced in order to pack two images into a single video frame. Therefore, the image detail provided by this format is moderate and less than what HDTV offers. In certain situations, even if the display is capable of showing much elaborated resolution (e.g., with 1080p active stereoscopic display), the frame-compatible format does not allow the full potential of the display to be utilized.

Apart from the abovementioned resolution issue, which is specific to the selected picture format, the stereoscopic concept, in general, has a number of limitations due

Table 14.1 Frame-compatible formats for 3DTV broadcasting

Format	Description	
Side-by-side		Left and right views are horizontally subsampled and stitched side by side
Top-bottom		Left and right views are vertically subsampled and stacked them vertically
Row interlaced		Left and right views are vertically subsampled and pixel rows from two resultant images are interlaced with each other
Column interlaced		Left and right views are horizontally subsampled and pixel columns from two resultant images are interlaced with each other
Checkerboard		Quincunx subsampling is applied to the left and right views before packing pixels from each view into a checkerboard pattern

to a number of visual artifacts such as depth nonlinearity, sheer distortion, and depth and size magnification [4]. Therefore, stereoscopic content itself is not capable of providing the natural 3D sensation. In order to resolve these issues, 3D video has to move forward overcoming the stereoscopic barrier. One of the most popular approaches is the multiview video together with depth maps or disparity maps, which is commonly identified as MVDx format, where x represents the number of views.

14.3 Beyond Stereoscopic Video: The MVDx Format

Multiview displays can produce different views depending on the viewing position. As a result, any head movement will result in viewers seeing different views of the same scene. Hence, more natural 3D perception can be reproduced avoiding artifacts such as sheer effect [5]. Besides, these multiview displays offer glasses-free 3D experience, unlike traditional stereoscopic displays. However, the down-side of multiview display concept is that they require content acquired from a large number of densely packed viewpoints. For example, a 64-view multiview display requires a video captured at 64 different viewpoints. This approach is prohibitively expensive in terms of capturing, storing, and distribution. Therefore, reduced number of views is the practical alternative, which compromises the view density. However, with the advancement of video processing technologies, it is now possible to generate novel views by interpolating real camera viewpoints to increase the number of viewpoints [6]. With this technology, the number of real viewpoints that should be captured can be minimized.

In order to reduce the complexity of the rendering algorithms, it has been proposed to exploit depth-image-based rendering (DIBR) for view interpolation. Although DIBR helps to reduce the cost of view renderer unit in a 3D display, it requires accurate depth or disparity maps to achieve artifact-free rendering. The format in which multiview content is provided with depth or disparity maps is called multiview-plus-depth/disparity (MVD) format. Often, the number of real viewpoints available in the 3D video is also identified with MVDx naming format. For example, 4-view video with disparity maps is an MVD4 format video.

Depth maps can easily be captured using depth-range cameras. However, this convenience comes with a cost. Under this approach, every camera has to be equipped with high-resolution depth-range camera. Such an approach is far from practical. An alternative is to derive disparity or depth information from the camera viewpoint images themselves. Even if the accuracy of depth/disparity map estimation is yet to become mature enough for practical applications, it is envisaged that the technology will help to simplify capturing process while still producing good-quality depth/disparity maps for multiview displays to operate at their best.

Besides novel-view generation, DIBR also helps to adapt the stereoscopic baseline to match viewers' interocular distance. Such adaptation not only helps to reproduce 3D scene accurately but also improves the viewing comfort. Furthermore, the scene geometry can also be adapted to display size in order to minimize the 3D distortion coursed by display size mismatches.

Since the MVDx format is foreseen as an acceptable compromise for the ever-increasing number of view density supported by 3D displays, a number of research initiatives have investigated multiview video capturing, depth/disparity estimation, postproduction, transmission, and display-rendering technologies to outline the future 3D broadcasting with MVDx format video. The MUSCADE project adopted MVD4 format which comprises of two stereoscopic inner view pair and two wide baseline satellite views as shown in Fig. 14.2. The satellite pair is about 60 cm apart for supporting wider baseline display technologies such as light-field displays.

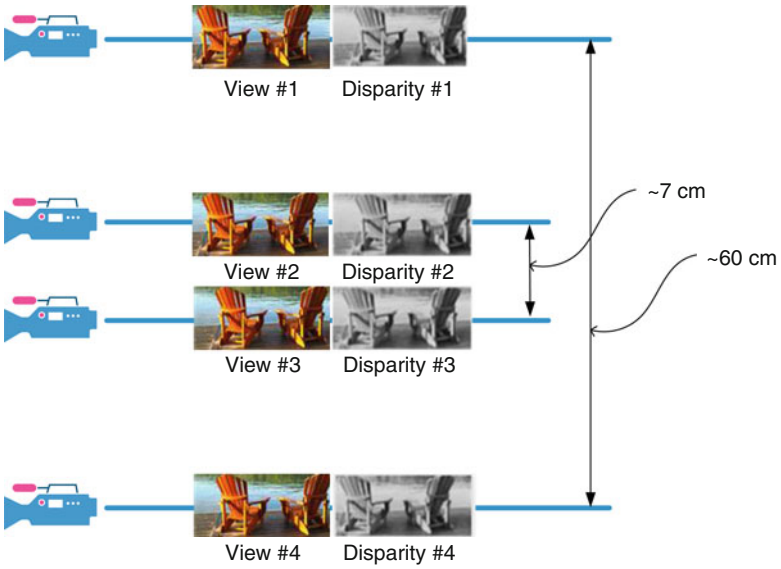


Fig. 14.2 MUSCADE MVD4 format

14.4 Spatial Audio for 3DTV

Compared to 3D video, spatial audio technologies are far more mature and welcomed by the home audience. The first known example of spatial audio is Clement Adler's 80-channel spatial telephone, which was demonstrated in 1881 [7]. At present, there are a range of spatial audio formats available within consumer domain. There are two basic spatial audio formats, channel-based audio and object-based audio. Channel-based audio is recorded with an array of directional microphones. A compatible surround sound reproduction system has a dedicated speaker to reproduce corresponding recorded channel. If these speakers are placed appropriately, the original audio scene can successfully be reproduced at any other location. The object-based audio format, on the other hand, records each sound source in a scene separately along with its location information. Compatible reproduction system uses location information for mixing individual sources together in order to correctly reproduce the original sound scene. The main advantage of object-based audio is that the recording format is independent of reproduction device. However, given an arbitrary and dynamic scene, object-based audio capturing is far more challenging than channel-based capturing. Moreover, home environment is more likely to be equipped with channel-based audio reproduction systems. Therefore, object-based audio has no significant advantage for consumer applications. Therefore, channel-based approach is opted into the MUSCADE project for delivering spatial audio.

14.5 3D Audiovisual Acquisition

14.5.1 Video Acquisition

Figure 14.3 illustrates a wide baseline video acquisition module, which supports four professional-grade cameras. Two central cameras of the proposed rig-mounted camera module are placed at stereoscopic baseline in order to ensure a backward compatibility with existing 3D broadcasting infrastructure. These two cameras are sandwiched between two additional cameras, which are called satellite cameras. In order to provide extra flexibility for various application scenarios, the camera rig is equipped with baseline adjustment facility. This feature is quite useful for attaining the full potentials of modern display types, such as light-field and multiview displays, which require wide baseline video. Furthermore, the camera rig also helps to precisely align all camera centers to a common center line. The inner camera pair has mounted in a parallel axis configuration in order to avoid keystone distortion. As a result, the convergence plane should be adjusted as a postproduction operation by shifting views towards one another. However, the satellite cameras have been mounted in such a way that convergence angle can be adjusted to ensure that the right amount of overlapping can be achieved amongst all camera viewpoints.

The greatest challenge with more than two camera setup is the aligning all cameras precisely to minimize unwanted vertical disparities. One of the roles of the assistance system is to simplify this cumbersome operation by guiding through

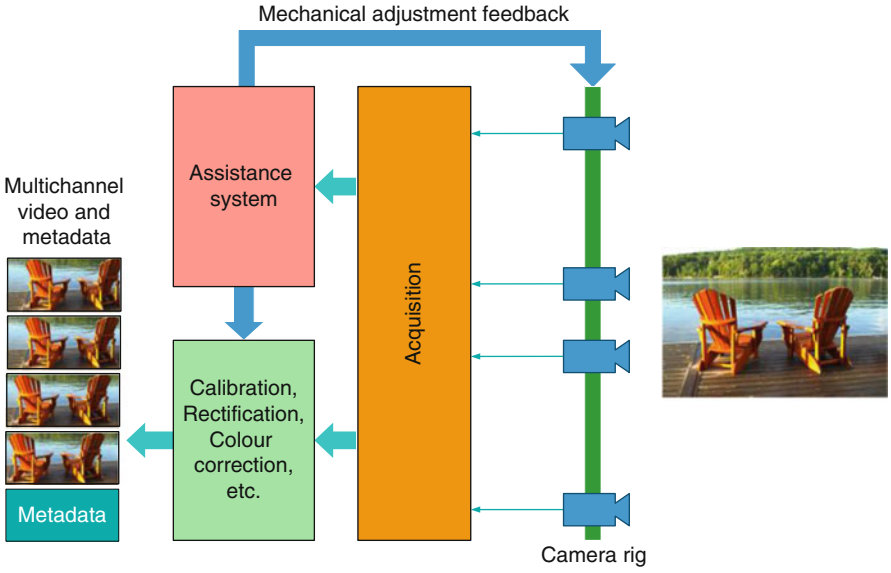


Fig. 14.3 Audiovisual acquisition module

the camera setting not only before the actual shooting begins but also during the shooting. This system analyses incoming video from cameras and provides the feedback to adjust camera geometric parameters such as position and orientation of camera. Moreover, the camera assistance system also makes sure whether basic rules of 3D acquisition are followed in order to avoid any viewing-comfort-related issues. In addition, it has been equipped with metadata-recording capabilities for securing data for color matching and view rectification operations in order to maintain homogeneity in color and geometry across all the acquired views.

Even though the camera rig is inevitably bulkier than 2D or stereoscopic camera rig, it has been successfully demonstrated under various indoor and outdoor environments under stationary as well as moving on track-and-dolly setup.

14.5.2 Audio Acquisition

At least one microphone array capturing audio in A or B format is required per scene. A format audio is captured with four capsules positioned at the sides of a tetrahedron. Small and compact A format microphones are easier to handle and can easily be hidden in the scene. It has no preferred direction of operation and, therefore, captures sound from every direction uniformly. Professional-grade A-format microphones have broadband operation (i.e., 20 Hz to 22 kHz) ability as required for broadcasting applications. Moreover, it generates only a small number of audio channels, which is an extra advantage for interfacing, storage, and processing. The signals captured by these capsules can be converted into B format, which consists of the pressure signal, W, and pressure gradients X, Y, and Z, along the corresponding axes. Professional-quality microphone arrays capturing in these formats are commercially available. These are conventionally used for ambisonics, 2-channel stereo, and 5.1 channel recordings.

14.6 3D Audiovisual Postproduction

On top of the traditional postproduction tools, a number of specialized tools are required. Figure 14.4 shows the audio/video postproduction module for offline and live postproduction. This module performs the following functions for preparing content captured with multiview camera setup.

View rectification: Multiview video rectification is more challenging than stereoscopic rectification. A common baseline is calculated considering all camera positions. Subsequently, the rectification transformation defined by camera rotations and camera intrinsic parameters is applied to all the views [8, 9].

Disparity/Depth map estimation: Accuracy of depth/disparity map has a profound impact on the quality of rendered views. Therefore, depth/disparity estimation is

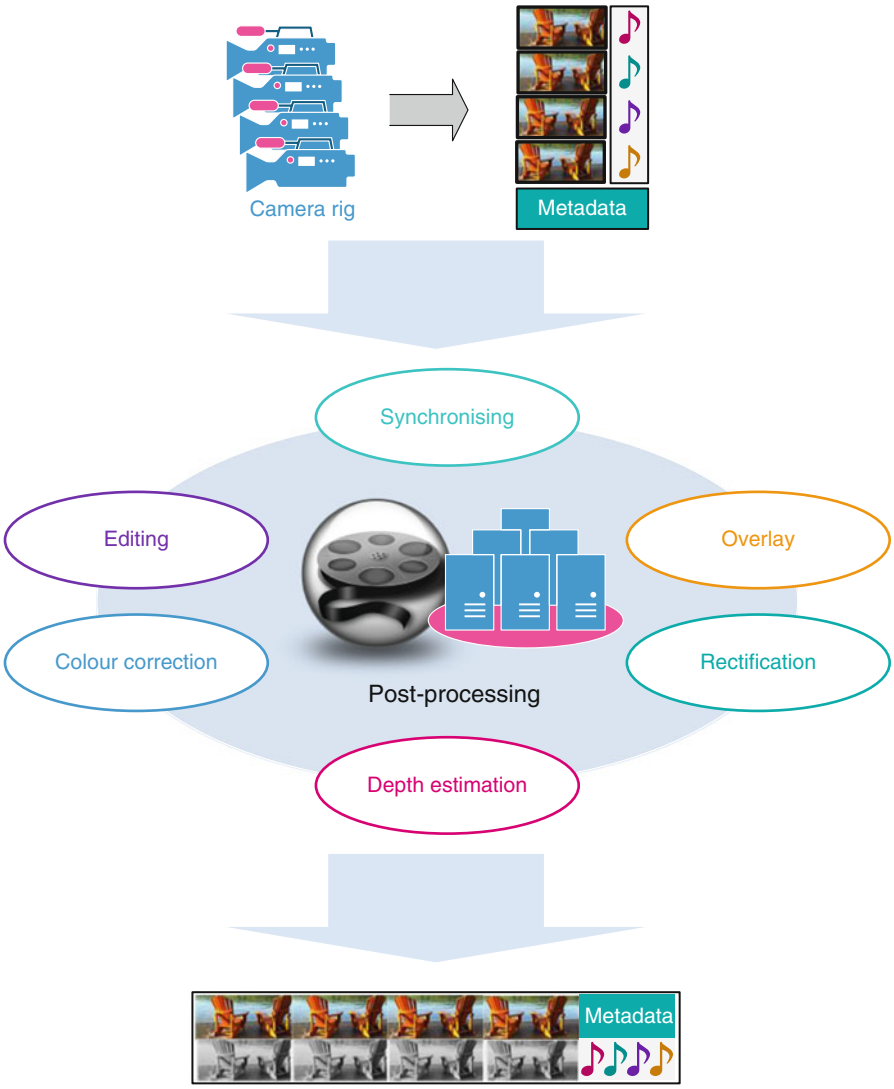


Fig. 14.4 Postproduction module

one of the vital function in MVDx content creation. Traditional disparity estimation algorithms, which mostly rely on simple block matching, are not ideal for depth/disparity map creation. In order to generate a good-quality depth/disparity maps, a combination of initial disparity estimation using the line-wise hybrid recursive matcher and a subsequent post-processing and up-sampling step using variations of cross-bilateral filtering have been proposed [10]. Real-time capability for image resolutions up to HD, scalability for large disparity ranges, and high degree of parallelization are important characteristics of these algorithms that are aimed at

broadcasting applications. Moreover, it is also necessary to make sure the depth/disparity maps are temporally consistent.

Color correction: Even though this is a traditional post-processing step commonly used in 2D applications, specific attention has to be payed for maintaining the color consistency across all the views. Any color offset amongst views has a detrimental impact on compression efficiency when inter-view prediction is used for encoding the multiview video. Therefore, color correction in multiview scenario often incorporates consistency optimization step along with the traditional algorithms [11].

Graphics overlays and subtitles: Traditional rules dictating the adding graphics overlays and subtitles are not always ideal for 3D scenes. Multiple concerns arise when subtitles and graphics overlays are added to a given scene. First of all, any object added during postproduction should not violate the geometric integrity of the scene [12, 13]. Any inconsistency will result in potential viewing discomfort. Furthermore, occlusion has to be correctly handled to maintain the naturalness of the scene. The other concern is the complication arising due to different displays requiring different formats and resolutions [14].

Editing: Due to additional dimension and irregularity of 3D video, editing is more challenging than 2D video editing. Even though basic operations such as trimming and rearranging clips seem to be straightforward, a product aiming at mass audience has to be carefully authored due to the notorious discomfort issue associated with 3D video viewing [15]. Factors that aggravate discomfort such as excessive disparity and rapid depth changes have to be carefully considered while editing 3D footages [16].

14.7 Encoding and Encapsulation

Multiview video and depth/disparity maps are encoded using multiview extension of H.264/AVC standard-compliant encoders. Figure 14.5 illustrates the encoding module. The layered structure provides scalability across view dimension. The main reason for choosing this encoding strategy is to support a large range of displays including legacy 2D, stereoscopic, multiview, and even advanced light-field display. The scalability structure guarantees that the same bit stream can be decoded by all kind of 3D displays. The stereoscopic base layer consists of two full HD-resolution frames. 2D support is provided through the inherent view scalability feature in multiview extension of H.264/AVC standard. A notable point in the encoding structure is that satellite views have been encoded independent of center stereoscopic view pair. This is because inter-view prediction between a view from center pair and another view from satellite pair does not produce any bit rate saving due to large separation between those views. Moreover, independent encoding also implies independent decodability. This is especially useful because it simplifies parallel decoding architectures. It should be noted that due to similar reason, inter-view prediction between satellite view pair is also not efficient. However, in order to maintain the consistency, these views are also encoded using a multiview encoder, which produces a single bit stream for the satellite view pair.

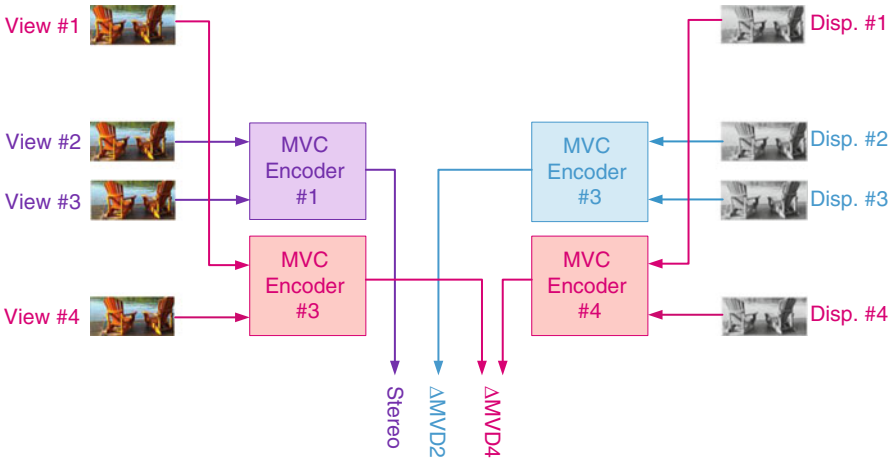


Fig. 14.5 Video encoding

The proposed scalability structure enables the decoder to drive the 2D and stereoscopic displays directly. However, view synthesizing stage is required for driving multiview and light-field displays. Moreover, the baseline adjustment [17] at the receiver is also possible by using stereoscopic base layer and ΔMVD2 layer. This feature is especially useful for personalizing the depth perception in order to minimize eyestrain [18]. Proposed baseline adjustment eliminates potential geometry distortion issue in widely used pixel shifting between the views.

An AAC encoder configured for multichannel encoding is used for encoding audio signal.

The encoded spatial audio, MVD4 video, and other information are encapsulated into three MPEG transport streams (TS). The first transport stream (TS#1) carries MVC encoded base layer video (i.e., stereoscopic video), AAC encoded multichannel audio, interactivity metadata, PMT, and PAT. MVC encoded disparity/depth maps associated with the stereoscopic pair (i.e., ΔMVD2 layer) are delivered over the second TS (TS#2), and the third TS (TS#3) carries MVC encoded satellite views and associated disparity/depth maps. Since TS#1 carries the most important information, it is expected to be delivered over a relatively more reliable channel. Synchronization amongst audio, video, and interactivity components is achieved through PTS time stamps. Moreover, MVD4 video has been encoded as four independent MVC stereo pairs (i.e., 2 view pairs and 2 disparity/depth pairs) and hence each stream has its own PID.

14.8 Distribution

Table 14.2 shows the data rates produced by MVD4 video and its potential sublayers. From the table, it is clear that MUSCADE content has a peak data rate of 37.2 Mbit/s. Together with audio and metadata, MUSCADE has to handle an

Table 14.2 Bandwidth requirements

Video format	Data rate (Mbit/s)
HD monoscopic 1080p25	8
MVD1 (color + depth/disparity)	12
Stereo (multiview encoded)	13.6
MVD2	20.4
MVD4	37.2
Simulcasting of 4 views + 4 depth/disparity maps	48

excess of 40 Mbit/s data rate altogether. Potentials of DVB and IP families of distribution networks were investigated under the MUSCADE project for broadcasting 3D video. This section summarizes the findings.

14.8.1 DVB Family Transmission Networks

The DVB family consists of four sister standards [19]. DVB-T standardizes the terrestrial broadcasting infrastructure, while DVB-S outlines the satellite-based digital video broadcasting. Cable infrastructure often distributes television channels using DVB-C standard. The handheld mobile terminals are the target platform for DVB-H standard. Out of these four standards, MUSCADE consortium concentrated on DVB-T- and DVB-S-based distribution links.

14.8.1.1 Satellite System

As shown in Fig. 14.6, typical satellite system consists of:

- A satellite offering a coverage over a particular service area through a single wide beam
- A gateway transmitting the 3D-TV programs to the satellite
- Terminals allowing the reception of 3D-TV programs at the end-user side
- A satellite control center that monitors the satellite's operations and rectifies any issues

The gateway transmits 3D television programs to the satellite (uplink) and hence it provides access to the satellite for the broadcasters. The DVB-S2 modulator at the gateway incorporates the scalability scheme defined for MVD4 content through Variable Coding and Modulation (VCM), which maps each scalability layer to a specific modulation and coding (MODCOD).

The satellite amplifies the signal and retransmits back to a wider coverage area. Due to higher bandwidth requirement, a Ku-band-based broadcasting is foreseen for 3D broadcasting. The Astra 1E satellite located at 23.5° east is an example of operational Ku-band-enabled satellite. The saturated EIRP performance over the satellite coverage area is shown in Fig. 14.7. This footage is in the Broadcasting

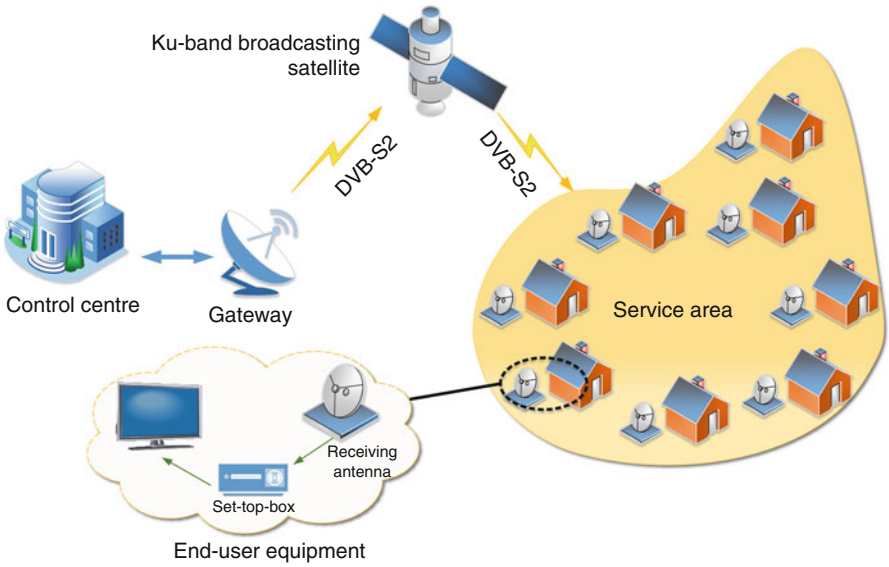


Fig. 14.6 Ku-band broadcast satellite system architecture

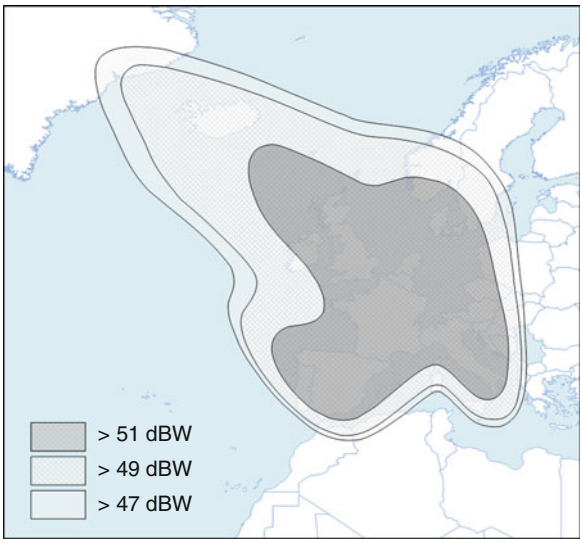


Fig. 14.7 Astra 1E EIRP performances (BSS, V-pol)

Satellite Service (BSS) frequency band (i.e., 11.70–12.10 GHz) and with vertical polarization. The transponders in BSS frequency band are of 33 MHz bandwidth and fed with a single 27.5 Mega symbols per second (Msps) carrier. Each carrier provides 42 Mbps with a 60 cm antenna in QPSK4/5 transmission mode and 53 Mbps with an 80 cm antenna in 8-PSK2/3 transmission mode.

The user terminal has an antenna and a set-top box. The parabolic reflector, which is often identified as the satellite dish, is the most common type of terminal antenna. The parabolic surface of the antenna reflects the signal to its focal point where the feed horn is placed. The feed horn converts electromagnetic signal into an electrical signal, which may be amplified using a low-noise amplifier. The size of the antenna may vary. Typical antenna sizes for home reception in Ku band range from 60 to 100 cm. Small antennas are deployed at the center of the coverage where satellite EIRP is high. The poor signal level at the edge of the coverage area is compensated with a larger antenna, which offers better reception gain. However, the use of a large antenna is not limited to an edge of the coverage areas. They can also be used at the center of the coverage to increase the link availability, which in turn increases the quality of service.

The set-top box supports VCM defined in DVB-S2 standard and consists of a video decoder and renderer. The decoder and renderer operate in a coordinated fashion in order to pick up the right scalability layer to support multiple display devices ranging from 2D to light field (Figs. 14.1 and 14.12).

Assuming that the service to be received is over 97 % of the coverage area and that the target time availability is higher than 99.7 %, the available bit rate is estimated to be 42 Mbps in the 60 cm antenna scenario and 53.2 Mbps in the 80 cm antenna scenario. Hence, a full transponder satellite channel can technically support MVD4 broadcasting. However, it is very unlikely that sustainable commercial service can be operated under such a high bandwidth, considering the cost of a transponder. However, under the current bit rate estimations for multiview 3D video, it is envisaged that MVD2 format will be more realistic until encoding algorithms are mature enough to compress MVD4 video into a commercially sustainable level.

14.8.1.2 Terrestrial System

The terrestrial broadcasting is operated through DVB-T/DVB-T2 digital terrestrial transmission (DTT) architecture shown in Fig. 14.8.

The DTT headend collects the television programs from the content providers for processing and distribution. The incoming TS streams are often processed at the headend to include additional A/V programs and auxiliary contents such as interactivity data, signalling, and EPG information. The resulting TS is subsequently delivered to the terrestrial transmitting sites through a dedicated distribution network. The distribution network is planned based on the number of transmitter sites. Point-to-point digital radio links, IP networks, optical fibers, or satellite transponders are often contracted for backhauling. In certain scenarios, a combination of technologies can also be used in parallel or sequentially. For example, the main transmission sites are served with a digital radio link in addition to a satellite link to increase the availability, while small/minor transmission facilities are served with the satellite link only.

Transmission sites receive television programs from the headend and broadcast through an omnidirectional antenna. The number of transmission sites required to cover the target service area varies from a few tens up to some hundreds, depending

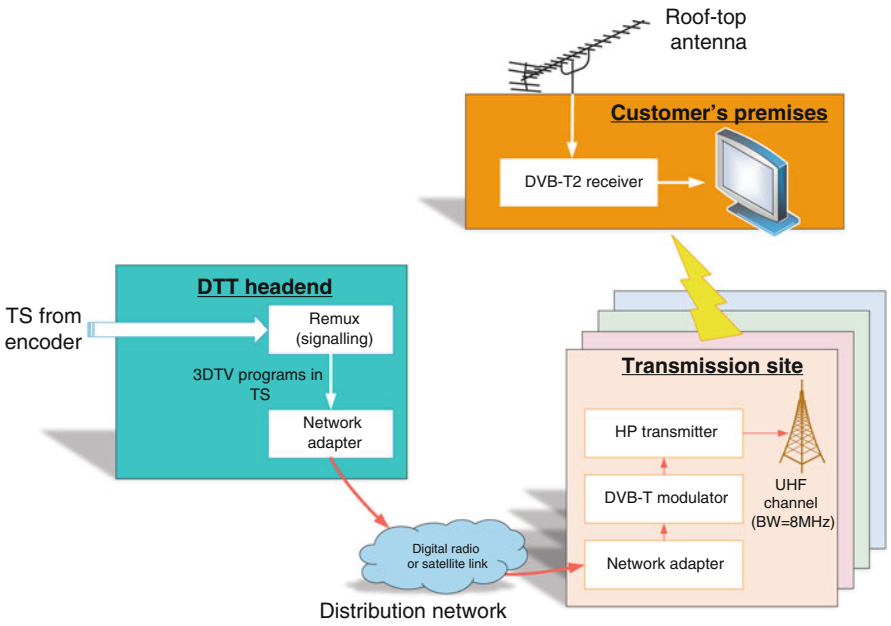


Fig. 14.8 DVB-T/DVB-T2 distribution channel

on the area to be covered and geographic condition. At the user premises, DVB-T or DVB-T2 signal is typically received through a rooftop antenna. This signal is decoded and played back by a set-top box that supports MUSCADE formats.

DTT services are operated using the same radio frequency bands allocated for analog television services. The frequency range allocated for a traditional analog UHF or VHF channel provides a typical data rate of 24 Mbit/s (DVB-T) or 40 Mbit/s (DVB-T2). Such a digital channel provides enough data rate for delivering a number of 2D television programs, which are multiplexed to a single stream identified as a multiplex. However, when it comes to MVD4 content, DVB-T2 channel capacity is barely enough for a single television program. However, DVB-T does not provide sufficient capacity for delivering MVD4 content at all. It can only carry the base (stereoscopic) and ΔMVD2 layers. Consequently, terrestrial broadcast channels are not yet commercially viable unless a drastic reduction of the data rate can be achieved through novel video encoding technologies.

14.8.2 IP Family

Two different network architectures were investigated for delivering MVD4 content under the MUSCADE project. They are:

- Wireline Internet Protocol television (IPTV) distribution
- Broadband Wireless Access Networks

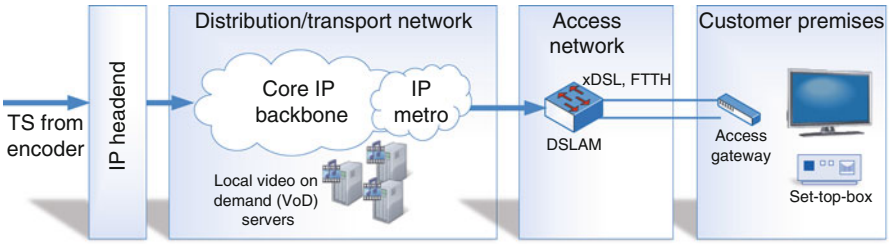


Fig. 14.9 Particular case of the IPTV wireline distribution

14.8.2.1 IPTV Distribution

IPTV over fixed line networks are delivered over managed networks. This option allows service providers to deliver vast amounts multicast video traffic as efficiently and effectively as possible. State-of-the-art managed networks supports MPEG-2 and H.264 in transport streams. Live television streams are usually delivered through IP multicast streams while Video on Demand (VoD) services are operated through IP unicast.

The IPTV broadcasting architecture is shown in Fig. 14.9. Similar to the DTT, the headend receives content from television content providers for processing and distributing over IP network. It hosts real-time encoders/transcoders, centralized video servers, media asset management system, middleware platform, conditional access (CAS), and DRM systems. The transport and distribution network relies on IP core network, which is usually an optical packet backbone or MAN GbE. The distribution network also hosts local VoD servers. The access network connects subscribers to their network service provider. Access network infrastructures use a range of technologies such as ADSL, ADSL2+, and optical access technology (FTTx). The home network is typically an Ethernet- and/or Wi-Fi-based network within customer's premises. This network hosts set-top box, TV, and other media equipment required for decoding and playing of 3DTV content.

The feasibility of delivering MVD4-based 3DTV services to home audience depends on the capacity of the access network, which is often the bottleneck in an IPTV delivery network. Access network infrastructures based on ADSL2+ technology [20] provides theoretical downlink capacity of 24 Mbit/s depending on the customer's location (especially the distance between the customer's premises and the Digital Subscriber Line Access Multiplexer (DSLAM)). Up to 12 Mbit/s bandwidth is typically reserved for IPTV service in a managed architecture. This capacity is more than sufficient for providing 2D HD television service. However, with present video compression technologies, this reserved capacity is not sufficient for at least the MUSCADE base layer (i.e., 1080p25 full frame stereoscopic video). Nevertheless, if theoretical maximum capacity of ADSL2+ access links is fully utilized, the base and ΔMVD2 layers can be delivered without any issue even though such scenario is neither practically nor commercially viable.

Present effort of telecom infrastructure companies to deploy FTTx optical access technology, in particular Fiber to the Home (FTTH), provides a glimpse of hope to

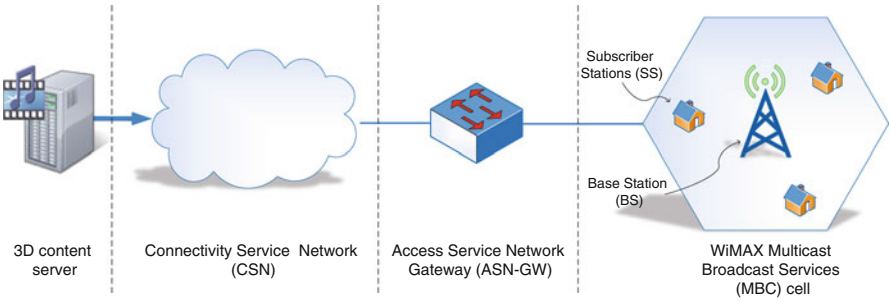


Fig. 14.10 Proposed 3DTV over WiMAX network architecture

provide MVD4-based 3DTV service over wireline IP networks. FTTH can provide a bandwidth of 100 Mbps to the end user. Therefore, MVD4 television service can be provided with a reserved capacity of about 40 Mbps. Moreover, scalable structure of MVD4 content as proposed in MUSCADE enables transmission and bandwidth occupancy optimization, taking into account bandwidth availability, which is time and location dependent.

14.8.2.2 Broadband Wireless Access Networks

WiMAX and WLAN architectures were investigated for delivering MVD4-based 3D content. Figure 14.10 illustrates the proposed network architecture for delivering 3DTV service to subscriber stations (SSs) over WiMAX. The connectivity service network (CSN) connects SS to the Internet. The CSN is owned by the network service provider (NSP). Access Service Network Gateway (ASN-GW) typically acts as a layer 2 traffic aggregation point. The base station (BS) provides the radio-dependent functions.

The WiMAX downlink provides a peak data rate of 27.216 Mbps using the maximum modulation and coding scheme (MCS) while keeping uplink to a downlink ratio of 1:0. However, this data rate is not fully available for the service since upper layer overheads (TS, RTP, UDP, IP, etc.) should also be accommodated within this data rate. Therefore, only the stereoscopic base and ΔMVD2 layers (19 Mbit/s in total) can be supported by the WiMAX service. However, in this case, the WiMAX base station effectively operates in a broadcast mode leaving little or no room for an uplink. Moreover, operating BS at the maximum MCS effectively limits the service area of the BS. Hence, the feasibility of WiMAX for 3DTV is highly questionable.

In contrast modern WLAN technologies supports far better data rate than WiMAX. Multi-gigabit data rates have been forecasted with upcoming technologies [21]. A typical WLAN architecture consists of an access point (AP) and one or more clients that connect to the AP through radio links as shown in Fig. 14.11. The AP acts as a gateway point between the wired network, which is commonly Internet enabled,

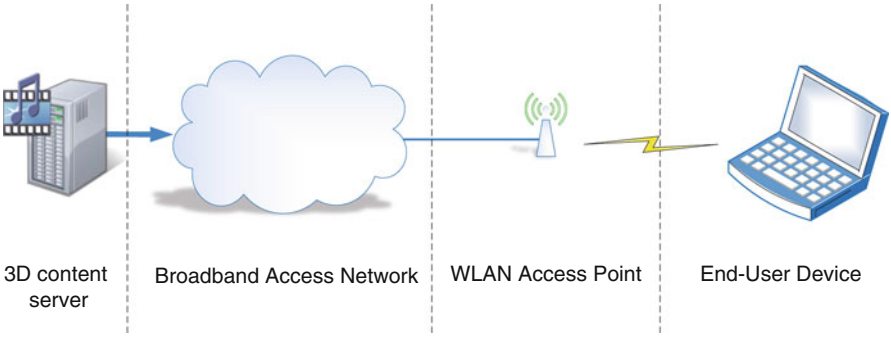


Fig. 14.11 Proposed 3DTV over WLAN network architecture

and the client. Since WLAN is an indoor end-user device, it can be used to channel IPTV content wirelessly over to a 3DTV set-top box connected to the home network.

A 15 Mbit/s minimum data rate (with 40 MHz bandwidth) supported by IEEE 802.11.n standard-compliant access points is barely sufficient for a stereoscopic baseline of MUSCADE 3DTV content. However, the data rate can go up to 150 Mbit/s, which can conveniently accommodate MVD4-based 3DTV content. Therefore, it is envisaged that WLAN will become a feasible medium for delivering MVD4 content to devices wirelessly connected to the home network.

14.9 3DTV Player

The 3DTV player implements decoding, rendering, interactivity, and display adaptation functionalities. Input for the player are the three transport streams defined in section entitled "Encoding and encapsulation" above. The high-level architecture of the audiovisual stream player is depicted in Fig. 14.12. The network interface supports satellite, terrestrial, and IP networks. The output from the network interface are the abovementioned three TS streams, which are de-encapsulated by the TS demux module to extract MVD4 content, audio streams, subtitle for interactivity, and other data. It also aligns elementary stream ensuring synchronization amongst different MVC encoded video and disparity streams. MVC decoder decodes elementary MVC streams to extract raw video and disparity maps. Four MVC decoders are operated in parallel to extract MVD4 content. The audio DLL performs audio decoding and rendering functions to generate multichannel audio and subsequently adapted for the room acoustic and speaker setup. The interactivity module uses hidden subtitles embedded in the TS streams for providing on-screen interactivity cues and handles interactivity through an integrated interactivity browser. The display-rendering modules use MVD4 video for generating image formats that are required for different types of displays. For instance, for light-field displays, the display adaptation module synthesizes the required number of viewpoints (typically 64, 128, etc.) and converts them into light-field module image format. Due to

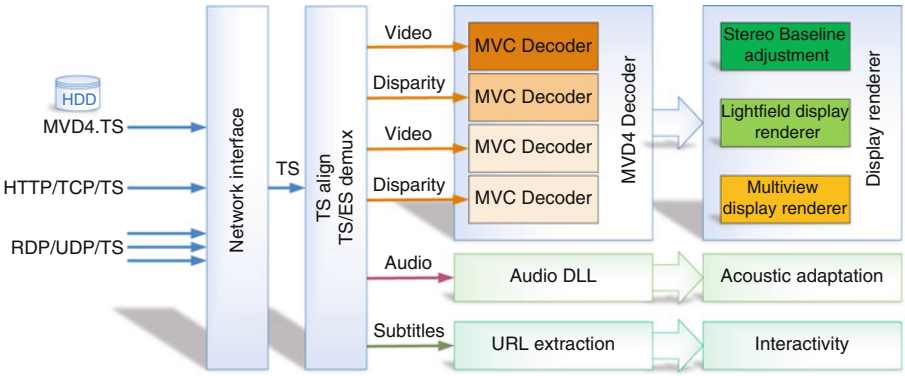


Fig. 14.12 Rendering, interactivity, and display modules

excessive computational resource requirement, modules of the audiovisual player have been distributed across a number of powerful processing hardware.

14.9.1 Interactivity Reference Architecture

The interactivity reference architecture proposed in the MUSCADE project distinctively identifies two domains as illustrated in Fig. 14.13.

14.9.1.1 Interactive Services Domain

The interactive services domain is made up of content provider, application platform provider, and network provider. The content provider creates audiovisual and interactive 3D content and links the latter into the former using metadata in the audiovisual stream. The linking process is also known as synchronization decoration since the audiovisual stream is decorated with correct associations between audiovisual and interactive content. The application platform provider publishes interactivity-enabled audiovisual content and distributes to the viewer when requested. This functionality is achieved through a 3D video interactive application (VIA) platform. VIA aggregates content and syndicate platform services with the interactive TV content provider across different delivery mechanisms (such as DTH, IPTV, and mobile). Content providers are provided with horizontal services and standard interfaces facilitating the development and deployment of video interactive applications. The VIA platform enables content to be either pushed to the viewer through broadcasting or pulled by the viewer on request to support content on demand. It also provides user management, application management, messaging, and payment services.

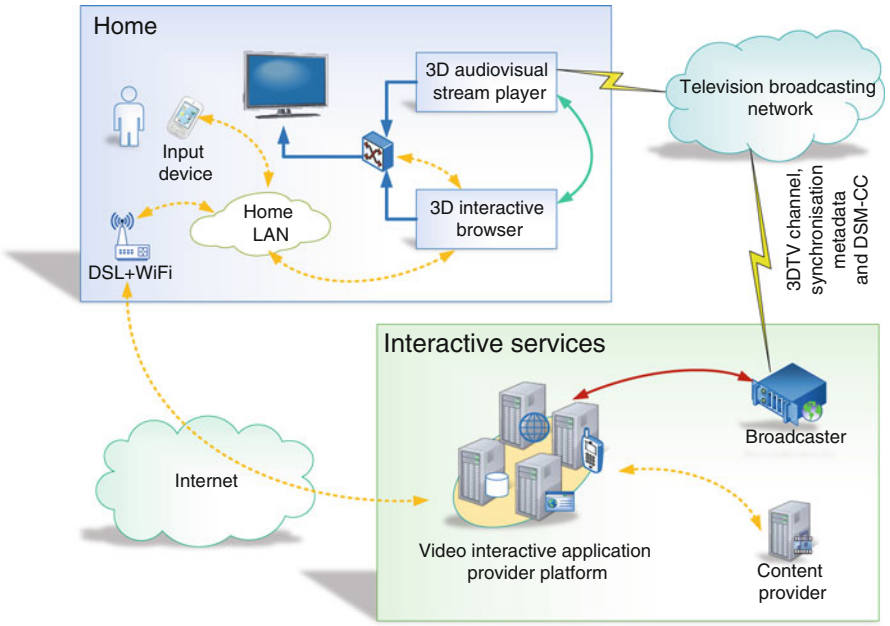


Fig. 14.13 Details of the interactivity architecture

The broadcaster or network provider operates the hybrid network through which the content is distributed to the viewer. Even though two-way communication is not essential for enabling interactivity, it enables providing more advanced forms of interactivity with 3D audiovisual content.

14.9.1.2 Home Domain

The home domain contains the 3D display, a soft-switch, the audiovisual player, the 3D interactive browser, an input device, and broadcast and return channel network access. The 3D display is the primary display on which the 3D audiovisual media is played. The audiovisual player decodes and renders audiovisual media on the primary display. It filters hidden subtitles in the media stream that notify interactive events and publishes them over the local network at appropriate time. The interactivity browser renders 3D interactive contents on the main display. It performs service discovery, user input handling, navigation functionalities, and game rendering. The switching between interactive browser and audiovisual player is handled by the soft-switch. This programmatically controls which source is shown on the 3D display. The interactive device is a secondary device that collects user input. It also acts as a display for personalized user interactions.

14.10 Conclusion

The MUSCADE project has successfully demonstrated an end-to-end delivery chain that supports MVD4 video, multichannel spatial audio, and interactivity content. Real-time content capturing, rectification, disparity estimation, postproduction, encoding, and encapsulation technologies developed under the MUSCADE project make it possible to produce MVD4-based live 3DTV content for wide baseline applications. The encapsulation architecture supports three levels of scalability for making 3DTV service available over a number of heterogeneous delivery infrastructures with a vast range of downlink capacities. Moreover, a 3D video player that supports real-time decoding and display rendering is also demonstrated successfully. An interactivity service platform is developed for interacting with 3D content on 3D displays. Feasibility of a number of transmission platforms has been assessed for MVD4-based 3DTV service provisioning. These include satellite (DVB-S2), terrestrial wireless (DVB-T2), IPTV, WiMAX, and WLAN. It is found that most of the present delivery networks do not provide a sufficient data rate to operate a commercially sustainable 3DTV service. This is mainly due to the video encoding technologies, which do not provide sufficient compression efficiency. Therefore, more research is required for developing compression technologies in order to deploy MVD4-based 3DTV services successfully.

Acknowledgments The work discussed in this chapter was performed within the MUSCADE integration project funded under the European Commission IST FP7 program.

References

1. Teulade V (2010) 3D Here and now... a goose that lays a golden egg? PricewaterhouseCoopers, Paris. http://www.pwc.com/en_GX/gx/entertainment-media/pdf/3d-technology-here-and-now-v2.pdf
2. MUSCADE–MUltimedia SCAlable 3D for Europe. http://www.muscade.eu. Accessed Jan 2010
3. Vetro A (2010) Frame compatible formats for 3D video distribution. In: IEEE International conference on image processing (ICIP), Nov 2010
4. Boev A, Hollosi D, Gotchev A. Classification of stereoscopic artefacts. MOBILE3DTV deliverable D5.1. [Online]. http://sp.cs.tut.fi/mobile3dtv/results/tech/D5.1_Mobile3DTV_v1.0.pdf.
5. Allison RS, Rogers BJ, Bradshaw MF (2003) Geometric and induced effects in binocular stereopsis and motion parallax. Vision Res 43:1879–1893
6. Zitnick C, Kang S, Uyttendaele M, Winder S, Szeliski R (2004) High-quality video view interpolation using a layered representation. In: ACM SIGGRAPH
7. Schoenherr SE (2001) Stereophonic Sound. [Online] http://www.aes.org/aeshc/docs/recording.technology.history/stereo.html
8. Kang Y-S, Lee C, Ho Y-S (2008) An efficient rectification algorithm for multi-view images in parallel camera array. In: 3DTV Conference: the true vision—capture, transmission and display of 3D video

9. Zilly F, Riechert C, Müller M, Waizenegger W, Sikora T, Kauff P (2012) Multi-camera rectification using line-arized trifocal tensor. In: International conference on pattern recognition (ICPR 2012), Tsukuba Science City, Japan, Nov 2012
10. Riechert C, Zilly F, Müller M, Kauff P (2012) Real-time disparity estimation using line-wise hybrid recursive matching and cross-bilateral median up-sampling. In: International conference on pattern recognition (ICPR 2012), Tsukuba Science City, Japan, Nov 2012
11. Doutre C, Nasiopoulos P (2009) Color correction of multiview video with average color as reference. In: IEEE international symposium on circuits and systems (ISCAS 2009), May 2009
12. Blonde L, Doyen D, Borel T (2010) 3D stereo rendering challenges and techniques. In: 44th annual conference on information sciences and systems (CISS), March 2010
13. Oh J, Sohn K (2012) A depth-aware character generator for 3DTV. IEEE Trans Broadcast 58 (4):523–532
14. Tourapis A, Fitzpatrick S (2012) Method for embedding subtitles and/or graphic overlays in a 3d or multi-view video data. US Patent EP2446636 A1, 2 May 2012
15. Lambooij M, IJsselsteijn W, Heynderickx I (2007) Visual discomfort in stereoscopic displays: a review. In: SPIE electronic imaging, stereoscopic displays and virtual reality systems XIV, vol 6490
16. Waschbüsch M, Würmlin S, Gross M (2006) Interactive 3D video editing. Visual Comput 22(9–11):631–641
17. Konrad J (1999) Enhancement of viewer comfort in stereoscopic viewing: parallax adjustment. In: SPIE stereoscopic displays virtual reality systems, vol 3639, pp 179–190, Jan 1999
18. Nojiri Y, Yamanoue H, Ide S, Yano S, Okana F (2006) Parallax distribution and visual comfort on stereoscopic HDTV. In: Proceedings of IBC, pp 373–380
19. Reimers U (2004) DVB—The family of international standards for digital video broadcasting. Springer, New York
20. ITU-T G.992.5, Asymmetric digital subscriber line 2 transceivers (ADSL2)—Extended bandwidth ADSL2 (ADSL2plus), Jan 2009
21. Wireless Gibabit Alliance (2010) Defining the future of multi-gigabit wireless communications. WiGig White Paper, July 2010

Index

A. Kondoz and T. Dagiuklas (eds.), *3D Future Internet Media*,
DOI 10.1007/978-1-4614-8373-1, © Springer Science+Business Media New York 2014