IPL

escola superior de tecnologia e gestão

instituto politécnico de leiria

DISSERTAÇÃO

Mestrado em Engenharia Electrotécnica - Telecomunicações

Subjective Quality Evaluation and Frame Loss Concealment in 3D Video

João Filipe Monteiro Carreira

Leiria, dezembro de 2012

escola superior de tecnologia e gestão
instituto politécnico de leiria

MASTER DISSERTATION

Electrical Engineering - Telecommunications

Subjective Quality Evaluation and Frame Loss Concealment in 3D Video

João Filipe Monteiro Carreira

Master dissertation performed under the guidance of Professors Pedro Amado António Assunção, Sérgio Manuel Maciel de Faria and Nuno Miguel Morais Rodrigues of Escola Superior de Tecnologia e Gestão of Instituto Politécnico de Leiria.

Leiria, December 2012

The true sign of intelligence
is not knowledge
but imagination.

(Albert Einstein)

Acknowledgments

I would like to thank to everyone that help during this research work, and that make it possible to accomplish.

I would like to express my gratitude to my advisers Prof. Pedro António Amado Assunção, Prof. Sérgio Manuel Maciel de Faria, and Prof. Nuno Miguel Morais Rodrigues, that in early times launched me to the research environment and trusted me during this journey. I'm thankful for their guidance and availability that were essential for the proper conduct of this research work. I also would like to thank their availability to review and improve my written English, in papers and this dissertation. I also would like to thank to the research group of P3DTV project, Prof. Rafael Caldeirinha, Prof. Telmo Fernandes and Tiago Figueirinha, for sharing their knowledge that help this research work.

I would like to thank the opportunity of working as researcher in the project ”P3DTV - Performance Optimisation in 3DTV Broadcasting Services” (P3DTV IT/IPLeiria/2009), for the important support to this work, and to thank Instituto de Telecomunicações and Escola Superior de Tecnologia e Gestão of IPL, for the laboratory facilities, that gave me conditions to accomplish this work.

I would like to acknowledge my research colleagues of Instituto de Telecomunicações - Leiria, Sylvain Marcelino, Lino Ferreira, Nelson Francisco, Luís Pinto, Luís Lucas, Anderson and Auridélia Moura, for the friendship and the wonderful work environment. I also want to express my gratitude for my graduation and non-graduation friends, for the moments outside the working environment, friendship and support, which were important along this research work.

A special thanks to my parents, António and Maria, and my sister Marta to whom I owe all that I have become. I also would like to express my gratitude to Bianca Pires for her support and love during this journey.

Finally, I would like to thank the participants of the subjective assessment trials performed during this research work.

Abstract

This dissertation presents a research work on frame loss error concealment techniques in 3D video.

An investigation on the impact of frame loss in the perceptual quality of 3D video is performed during this research work. The effect of temporal distortion due to frame loss in 3D video was subjectively evaluated. Various frame loss concealment methods were tested in order to evaluate their influence on the perceived 3D quality. It is assumed an error-free 2D video service, by using a high priority transport channel, thus, only the auxiliary view is affected by frame loss. Results show that the subjective 3D quality depends on the concealment method and the disparity of the original sequence. Another relevant result achieved is the evidence that it is perceptually better to switch from 3D to 2D viewing, rather than conceal 3D video, under heavy frame loss conditions.

A simulation study of a 3DTV broadcasting system over hierarchical DVB-T, based on the H.264/MVC, is presented. A 2D system is ensured by transmitting the base view over the high priority channel, while the auxiliary view is carried on the low priority channel, and subject to errors. Different quality models are proposed to evaluate the quality of the auxiliary view, using two well-known quality metrics (*i.e.*, PSNR and SSIM) and a recently proposed metric to evaluate the stereo perception. Since the results show that the proposed models present a good fidelity with the experimental data, they provide relevant insight to 3DTV network planning using the hierarchical DVB-T mode.

Efficient error concealment techniques are proposed, for 3D video decoders capable of handling stereoscopic views compliant with H.264/MVC standard. A joint motion-disparity compensation method is proposed to fully recover an estimated motion field for the lost frame in stereoscopic video, by using a combination of inter-view disparity-compensated motion vectors and intra-view motion extrapolation. Two methods based on this paradigm are proposed, and a decision method to find the more suitable method for different stereoscopic sequences The results show that the proposed method outperforms currently used methods such as frame-copy and motion-copy, and that the decision method is able to choose the more suitable of the two concealment strategies.

Keywords: 3D video, frame loss error concealment, subjective quality assessment.

Resumo

Esta dissertação apresenta um trabalho de investigação relacionado com técnicas de cancelamento de erros em vídeo 3D.

Neste trabalho é realizada uma investigação sobre o impacto da perda de imagens na qualidade perceptual do vídeo 3D. O efeito da distorção temporal devido à perda de imagens é analisado subjectivamente. Foram testados diferentes métodos de cancelamento de erros ao nível da *frame*, com o objectivo de avaliar a sua influência na qualidade da percepção tridimensional. Neste trabalho é assumido a existência de um serviço de vídeo 2D sem erros. Desta forma, apenas a vista auxiliar do vídeo 3D está sujeita a perda de *frames*. Os resultados obtidos demonstram que a qualidade subjectiva do vídeo 3D depende do método de cancelamento de erros aplicado, e também da disparidade das sequências de teste originais. Outro resultado relevante atingido é a evidência de que é perceptualmente melhor mudar de uma visualização 3D para 2D, em vez de corrigir os erros, quando ocorrem perdas excessivas de *frames*.

É apresentado um estudo de simulação realizado a um sistema de transmissão de TV3D sob DVB-T hierárquico, baseado na norma H.264/MVC. O vídeo 2D é assegurado pela transmissão da vista base usando o canal de alta prioridade, enquanto a vista auxiliar é transmitida pelo canal de baixa prioridade, e assim está sujeita a perdas. São propostos diferentes modelos para avaliar a qualidade da vista auxiliar, usando duas métricas bastante usadas (PSNR e SSIM) e uma métrica recentemente proposta para avaliar a percepção estereoscópica. Visto que os resultados obtidos mostram que os modelos propostos apresentam resultados semelhantes aos dados experimentais, eles são um contributo relevante para o planeamento de um sistema de televisão 3D baseado no DVB-T hierárquico.

Por fim são propostos nesta dissertação técnicas eficientes de cancelamento de erros para descodificadores de vídeo 3D compatíveis com o H.264/MVC. É proposto a combinação de compensação de movimento e disparidade para recuperar o campo de vectores da *frame* perdida na vista auxiliar. São propostos dois métodos baseados neste paradigma e um método de decisão. Os resultados obtidos mostram que os método propostos alcançam melhores resultados que o método *frame-copy* e o método *motion-copy*, e que o método de decisão é capaz de escolher entre o melhor dos dois métodos.

Palavras chave: Video 3D, cancelamento de erros, avaliação subjectiva de vídeo 3D.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-Dimensional, p. 1
3DTV	Three-Dimensional Television, p. 7
3D	Three-Dimensional, p. 1
ACR	Absolute Category Rating, p. 54
AD	Absolute Difference, p. 27
ATD	Absolute Temporal Difference, p. 85
BMA	Boundary Match Algorithms, p. 33
CIR	Concealed Image Ratio, p. 93
CSV	Conventional Stereo Video, p. 8
DCT	Discrete Cosine Transform, p. 30
DIBR	Depth-Image-Based Rendering, p. 11
DMVE	Decoder Motion Vector Estimation, p. 34
DPB	Decoded Picture Buffer, p. 18
DTS	Decoded Time Stamp, p. 67
DVB	Digital Video Broadcast, p. 2
DV	Disparity Vector, p. 40
ES	Edge Strength, p. 25
FC	Frame-Copy, p. 32
FLR	Frame Loss Ratio, p. 58
FPA	Frame Packing Arrangement, p. 10

GOP	Group Of Pictures, p. 18
GOP	Group of Pictures, p. 9
HD	High-Definition, p. 1
HMVE	Hybrid Motion Vector Extrapolation, p. 38
HVS	Human Visual System, p. 48
ICMV	Interpolated Candidate Motion Vectors, p. 34
ICV	Illumination Compensation Value , p. 45
IDR	Instantaneous Decoding Refresh, p. 17
ITU	International Telecommunication Union, p. 54
LCD	Liquid Crystal Display, p. 54
LDV	Layered Depth Video, p. 13
MBAFF	Macroblock-Adaptive Frame-Field, p. 19
MB	Macroblock, p. 23
MC	Motion-Copy, p. 32
MMA	Multi-frame Motion vector Averaging, p. 35
MOS	Mean Opinion Score, p. 55
MPEG	Moving Picture Experts Group, p. 1
MVC	Multi-View Coding, p. 4
MVD	Multi-View plus Depth, p. 12
MVE	Motion Vector Extrapolation, p. 37
MVP	Multi-View Profile, p. 14
MV	Motion Vector, p. 23
NAL	Network Abstraction Layer, p. 16
NCC	Normalised Cross Correlation, p. 45
OBMA	Outer Boundary Matching Algorithm, p. 34

P2P	Peer-to-Peer, p. 2
PES	Packetized Elementary Stream, p. 20
PMVE	Pixel-based Motion Vector Extrapolation, p. 37
POCS	Projection Onto Convex Sets, p. 30
PSNR	Peak Signal-to-Noise Ratio, p. 69
PVD	Preferred Viewing Distance, p. 54
QP	Quantisation Parameter, p. 86
QoE	Quality of Experience, p. 2
RMSE	Root Mean Square Error, p. 71
RTP	Real-time Transport Protocol, p. 2
SDV	Spatial Direction Vector, p. 25
SEI	Supplementary Enhancement Information, p. 10
SPS	Sequence Parameter Set, p. 15
SSM	Stereo Sense Metric, p. 4
SVC	Scalable Video Coding, p. 16
TS	Transport Stream, p. 20

Chapter 1

Introduction

This chapter presents an introduction to the research work carried out in the scope of this dissertation. The motivation and the objectives of the research are presented and the structure of the thesis is described.

1.1 Context and motivation

The developments in video coding over the recent years made possible the deployment of many applications over the internet such as video conferencing, video streaming, video publishing, among others. Also, demands for high quality multimedia resulted in new standards such as High-Definition (HD) video and High-Definition Television (HDTV) at higher resolutions, increasing the users' experience, pushing the technologies forward and opening new research directions.

The current trend in multimedia services and applications is to extend the sensory effect through new perceptual experiences, by including the additional dimension of depth in video content. Thus, nowadays, three dimensional (3D) video content and applications are emerging in the consumer market, broadcasting services and internet, as an extension of their existing 2D video counterparts, enhancing the user experience into a more immersive and natural viewing experience. The recent success of 3D movies and video games is seen as being partially responsible for the increasing interest in 3D visual contents, electronic devices, services and applications. Furthermore, there is also a rapid development of 3D video-based technologies for acquisition, processing, coding and displaying [1], allowing 3D video delivery services to be available to the general public through current communication networks, either at home [2] or in a mobile environment [3]. Associated with the 3D video there are other applications emerging, *e.g.*, the free viewpoint TV, which requires an higher number of viewpoints, requiring higher bitrate allocation in the transmission system. Another important application of multi-view video is the immersive

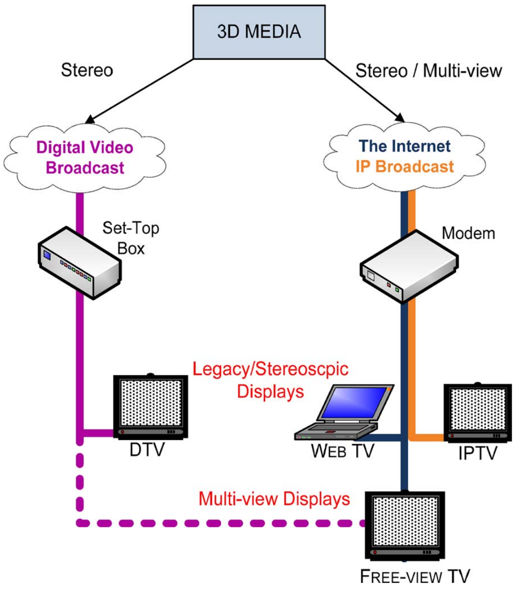


Figure 1.1: Platforms for 3D multimedia transport [10].

teleconferences. Beyond the advantages provided by 3D displays, it has been noticed that a teleconference system could enable a more realistic communication when eye contact is provided [4]. This requires an efficient compression of the video information, in order to be delivered under common 2D video channel without significant modifications. Thus, the recent developments in the field of 3D video coding and transmission are mostly based on either the H.264/MVC standard [5] or the video plus depth format, enabled by the MPEG extension known as MPEG-C Part 3 [6].

In the past years several 3D broadcast experiments were performed using different network topologies [7]. As for the current 2D video systems, the forthcoming 3D video services are also expected to be delivered over bandwidth constrained networks, comprised of error prone channels and heterogeneous transmission technology. As shown in Figure 1.1, different networks environments can be used to broadcast the 3D video information to the final users, covering from the IP-based, using the Real-time Transport Protocol (RTP) [8], to the terrestrial DVB networks [9]. Over this networks different techniques may be applied, such as asymmetric stereoscopic video streaming, adaptive peer-to-peer (P2P) streaming and view-selective streaming [10].

Despite the different characteristics of the existing 3D video formats and their associated compression techniques [11], the quality of the user experience (QoE) provided by delivery services, consistently depends on the need for adaptation of compressed streams to the diverse networking technologies, as well as to the transmission errors and/or packet losses [12]. As mentioned before, in generic multi-view video communications, the state-of-the-art codec H.264/MVC is currently used in most stereoscopic video applications. Since this is a highly predictive coding format, the MVC compressed streams are very sensitive to transmission errors and packet losses [13].

The perceived video quality is of highest importance for the adoption of a new technology from users' point of view, and consequently, this is also a critical market factor that is taken into account from the industry perspective [14]. Besides the already known distortion effects investigated in the past for the 2D video, *e.g.*, blockiness and blurriness, in 3D video there are new types of distortion arising from representation, coding, transmission errors and display, such as absence of depth clues, vergence and accommodation problems, and binocular rivalry. Thus, the existence of efficient error protection and error concealment techniques is a demand for the success of emerging 3D/multi-view systems. Since the increasing quality requirements conflict with bandwidth constrains, the availability of the 3D video services may be compromised. Therefore, an intermediate operational region may exist, where 3D service may not be available but 2D video quality is guaranteed.

Summarising, it is necessary to investigate the relevant subjective factors related with the 3D video quality in the presence of transmission errors, in order to find out where the quality of service meets the requirements of the 3D video applications. Moreover, efficient concealment techniques should be developed to cope with errors in 3D video, in order to smooth the effects of transmission errors.

1.2 Objectives

Due to the increasing number of services and bandwidth constrains, the quality of service may not be always guaranteed in 3D video. Therefore, the research community investigated for several years different concealment methods to reduce the impact of the transmission errors, which lead to video data loss and, consequently, to quality degradation.

Although several techniques have been proposed to recover the missing information in 2D video and more recently also for 3D video, the concealment of errors in 3D video is still an open research topic. In this context, this research deals with subjective quality evaluation and frame loss concealment of 3D video, with the following objectives:

- **Study the subjective and objective impact of frame loss in stereo video.** Since the impacts of transmission errors in 3D video are still an open issue, one of the objectives of this research work is to find impact factors of the temporal distortion caused by frame loss in stereoscopic video. Moreover, the influence of different concealment methods, which lead to different artefacts, should be tested, in order to assess the performance of the concealment methods in the perceived quality.

- **Evaluate the objective quality of the 3D transmission over hierarchical networks.** As mentioned before, the transmission channel may influence the quality of the 3D video streaming, therefore, it is necessary to evaluate the quality of 3D video transmitted over error-prone networks. In this work it is considered the use of a hierarchical channel where the quality of 2D video may be guaranteed. A simulation study of the transmission system should be performed, in order to devise quality models to predict the quality of the 3D video under different conditions.

- **Develop an efficient frame loss concealment method for 3D video.** Error concealment in 3D video is still under investigation by the research community. The final objective of this research work is to evaluate the performance of different concealment methods for the H.264/MVC decoder, and develop an efficient alternative for the existent methods, exploiting both temporal and spatial redundancies of the stereoscopic video.

1.3 An outline of this thesis

This thesis contains seven chapters and one appendix. The following chapters are organised as follows.

The first two chapters review the state-of-the-art concepts related with this research work. Chapter 2 describes the main technologies used in 3D video. The main representation formats of 3D information are presented. A description of the compression techniques for stereoscopic video is detailed, introducing the multi-view extension of the H.264 standard (MVC). Finally, some concepts associated with storage and transmission of 3D video are presented, in the context of 3D compressed video. Chapter 3 reviews some of the concealment techniques proposed over the last years, since they provide relevant knowledge to the main objectives of this work.

Chapter 4 presents a subjective evaluation study in order to find the relevant factors that influence the perceived 3D video quality. Three simple frame loss concealment methods are tested, and two sets of results are presented, corresponding to quite different frame loss conditions.

Chapter 5 describes different empirical models to predict the quality of the 3D video (stereoscopic video) using two well-known quality metrics (*i.e.*, PSNR and SSIM), and a recently proposed metric based on the disparity information, stereo sense metric (SSM). The models aim to predict the quality of the 3D video transmitted through a hierarchical DVB-T channel, for which, the left and right view of the stereoscopic video are delivered using different priorities. The framework of the broadcast system and the empirical model are described in detail.

Chapter 6 presents a contribution of this research work in the context of frame loss concealment. Different concealment schemes are described and tested using the Stereo High profile of the H.264/MVC codec, and the performance evaluation of each one is discussed. Then, two novel joint concealment techniques are proposed, in order to achieve an efficient reconstruction of missing frames under different conditions. The results show that the proposed method outperforms currently used methods such as frame-copy and motion-copy. Finally, a motion based decision to decide the best joint method is described, showing good performance under different stereoscopic sequences.

Chapter 7 concludes this dissertation and presents some suggestions for future work.

Appendix A presents the stereoscopic sequences used in the experimental study. Appendix B presents the published paper, which includes some results and conclusions presented in this dissertation.

Chapter 2

3D Video formats for coding and transmission

3D video content and applications are emerging in the consumer market, broadcasting services and internet, enhancing the users' experience into a more immersive and natural viewing experience. The rapid evolution of 3D technology spans from equipment (*e.g.*, stereo cameras and displays) to standards (*e.g.*, H.264/MVC [5]) and transport networks [15]. In current broadcasting services, 3D video is already reaching the home users. Major drivers for such new services are the frame compatible formats, which are being used to deliver 3D video services with minor changes in the standard compliant technology, providing a fast and economic way to deploy this new service [2]. In the near future, the multi-view coding extension of H.264/AVC will allow even more flexibility in these services, ranging from IP to digital terrestrial broadcasting, without losing backward compatibility to current 2D video systems.

This chapter reviews some of the main technologies of 3D video coding and transmission systems. The next section describes the main formats for 3D video applications: the stereoscopic video, which is the focus of this research work, and the depth-enhanced formats. Section 2.2 introduces some of the key features of the MVC extension of the H.264/AVC standard, as it is the state-of-art codec for the stereo and multi-view video. In Section 2.3 some approaches to transport and delivery of 3D services are presented. Finally, Section 2.4 concludes this chapter.

2.1 3D video: representation and coding formats

In the recent past, different formats for 3D television broadcast have been explored by the research community and related industry [16–18]. In this section the main 3D coding formats are described, introducing different 3D video data representations associated with

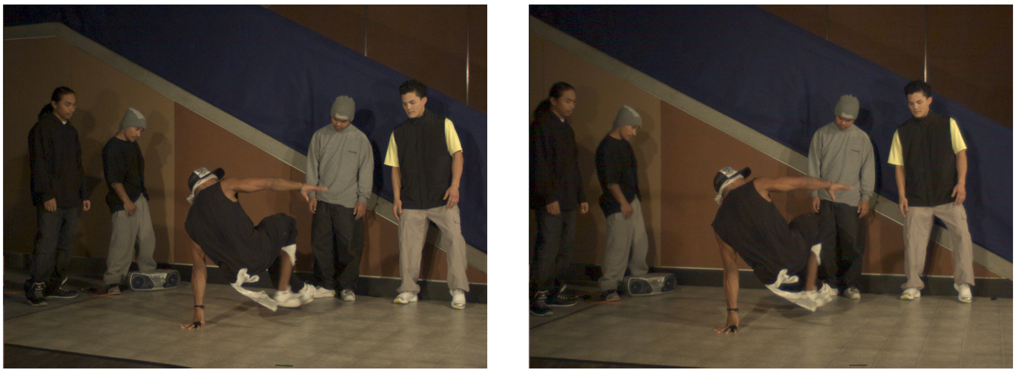


Figure 2.1: Representation of a 3D video scene using two views.

the 3D video applications, *e.g.*, 3DTV, 3D Blu-ray and FTV. These different 3D video data representations have different advantages and drawbacks, with regard to the efficiency, functionality and complexity according to the following key requirements: utilise existing delivery infrastructure, backward compatibility, minimal changes in the receiver components, support for a wide range of displays, and high 3D video quality.

There are different data formats for representation of a 3D scenario, such as: using stereo images from different cameras [19] and using an image with its associated depth, *i.e.*, video plus depth. For each representation, different coding formats may be used to exploit the redundancy between the 3D information and to maintain the compatibility with the current video systems.

2.1.1 Stereoscopic video

The stereoscopic video, also known as conventional stereo video (CSV), is a well-known and simple format for 3D video data representation [17]. A stereoscopic video can provide the 3D perception by using left and right views as a stereo pair. A stereo camera system simultaneously captures two views of the same scene to acquire video from slight different viewpoints. As a result, a pair of 2D views, as illustrated in Figure 2.1, is acquired corresponding to the views seen by the human eyes, one for the left eye, and the other for the right eye. The main drawback of the stereoscopic video format is the non-existence of multiple viewpoints, which does not allow motion parallax [20]. Nevertheless, in the current stage of the 3D video systems, most of the 3D applications for real world perception rely on this simple format.

Figure 2.2 represents three different approaches for coding and transmitting 3D content. All of theme are described in this section. The simpler coding approach for the CSV format is the simulcast. Figure 2.2(a) illustrates the case of simulcast coding and transmission. As shown, each of the two input sequences is independently encoded with the H.264/AVC [22–24], resulting in two independently compressed bitstreams, without

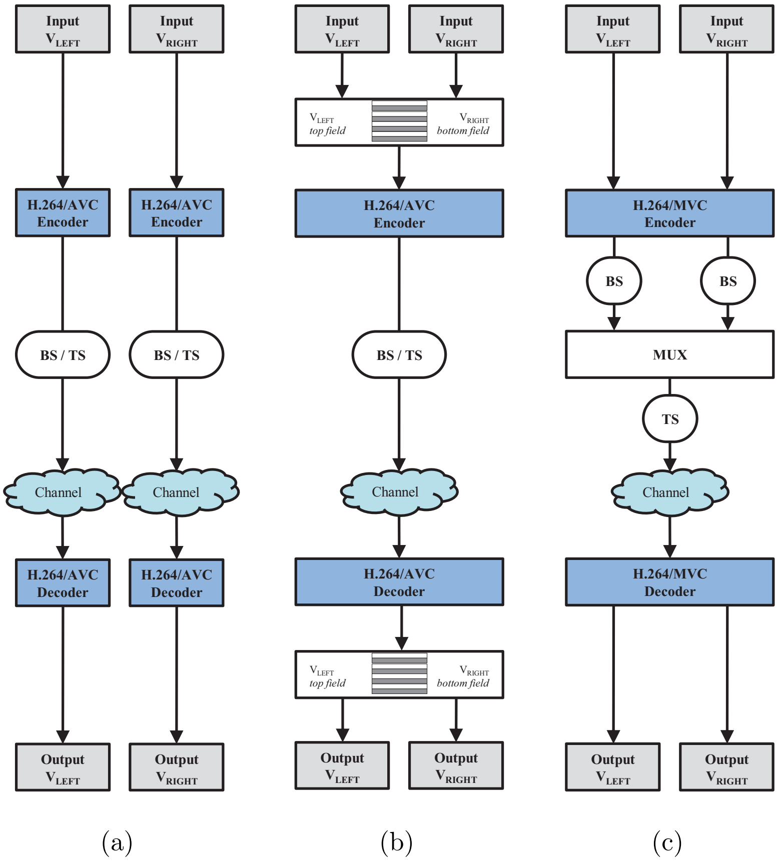


Figure 2.2: Schematic block diagrams for H.264/AVC Simulcast (a), H.264/AVC Stereo SEI Message (b) and H.264/MVC (c) coding with stereo video format data [21].

extra syntax elements. This format does not benefit from the redundancy between views (video signals), which results in high bitrate. The amount of data is proportional to the number of encoded views. However, the complexity of the transmission system remains the same of 2DTV [16]. The error propagation does not affect multiple views and the random access control between views only depends on the GOP's length.

As an alternative to the independent left-right views representation, frame-compatible formats [25] provide another type of 3D representation for storage and transmission of stereoscopic video. A common frame-compatible format is obtained by multiplexing the left and right views into a single composite frame and then into a sequence of such frames. This representation facilitates the distribution of the 3D/stereo video services using the existing infrastructure and transmission equipment already deployed for 2D video. Multiplexing of two views is performed using either temporal or spatial interleaving, and usually a sub-sampling operation is applied, in order to maintain the backward compatibility with the resolution of previous systems. A quincunx sampling may be applied to each view, so the interleaving of two views is performed with alternating samples in both horizontal and vertical dimensions [26]. Alternatively, the two views can be positioned in

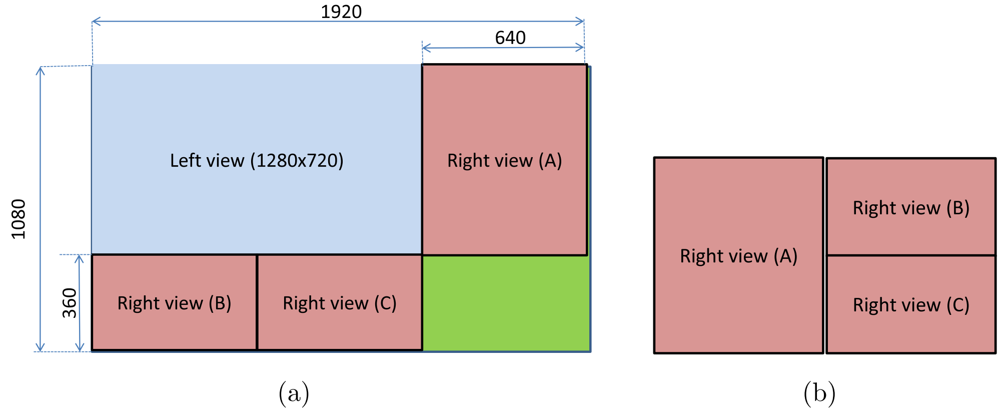


Figure 2.3: Illustration of the frame-compatible tile format: (a) final tile format comprising the left and right views; (b) tiling lines of the right view frame [27].

a side-by side or top-bottom format using a horizontal or vertical decimation respectively. Temporal multiplexing is also possible, where the left and right views are interleaved as alternating frames, with reduced frame rates, so that the total amount of data remains the same as that of a single view. A tile frame-compatible, as presented in the Figure 2.3, is also possible to use [27]. The figure illustrates the multiplexing of two views, with a spatial resolution of 1280 × 720, into a single composite frame of 1920 × 1080 pixels. Since this method does not need to reduce the spatial resolution, it is expected to achieve higher quality. Moreover, legacy 2D video systems just need to perform a simple crop operation to discard the right view information. The main drawback of this method is the artefacts around the tiling lines, however they are only noticeable for low bitrates. Frame-compatible formats have received considerable attention from the industry since they can be quickly deployed to home users with 3D capable displays. The main drawback of this representation format is the loss of spatial or temporal resolution, which may affect the quality of 3D perception. Moreover, view interleaving in standard streams needs additional signalling to be correctly de-interleaved and displayed.

The signalling for the frame-compatible formats was standardised within the H.264/AVC standard as Supplementary Enhancement Information (SEI). An example using SEI messages to transmit the stereos video is illustrated in the Figure 2.2(b). In general, SEI messages are used at the decoder to identify the left and the right views, but are not a normative part of the decoding process to reconstruct the pixel samples. The stereo SEI message, presented in a earlier version of the standard, has the capability of indicating whether the encoding of a particular view is self-contained, *i.e.*, the left view is only predicted using other frames or fields from the left view. Therefore, inter-view prediction for stereo is possible, by disabling the self-contained flag. Recently, a new SEI message referred to as the frame packing arrangement (FPA) SEI message was specified as an amendment to the H.264/AVC [28]. This new SEI message combines the above functionalities with additional signalling for various spatially/temporal multiplexed formats.

The previous coding and transmission methods for stereo video, namely the simulcast and frame-compatible formats, are simple approaches to deliver the 3D video services within the current technologies used for traditional 2D video. Although frame-compatible formats can exploit more spatial redundancies than simulcast, they cannot exploit those redundancies very effectively, since there is not a direct prediction of the right view from the left view. The approach represented on the Figure 2.2(c) uses the multi-view coding (MVC) extension of the H.264/AVC codec [29]. A key feature of the MVC design is the inter-view prediction [30], which enables a flexible use of reference picture management capabilities, already existing in H.264/AVC, by adding the decoded frames from other views for inter-picture prediction. This permits a more efficient reduction of the temporal and spatial redundancy, by exploring the similarities between nearby viewpoints. According to the MVC standard, inter-view prediction is constrained to the pictures contained within the same access unit, *i.e.*, pictures within to the same display instant. This reduces the coding complexity without relinquishing coding efficiency. Previous studies [18, 21] show that the multi-view coding can archive up to 35% more compression in comparison with simulcast, for the same objective quality. This reveals the relevance of inter-view prediction. As shown in the Figure 2.2, using the H.264/MVC the stereo video is encoded together, however it can be multiplexed separately allowing more broadcast flexibility. Further description of the key features of the MVC extension is presented in the Section 2.2.

2.1.2 Depth-based representations

Another representation approach for image-based stereo video is the video plus depth (V+D) [31]. Initially studied in the computer vision field, the video plus depth format provides a regular 2D scene representation enriched with its associated depth information, as shown in the Figure 2.4. The 2D video provides the texture information, *i.e.*, the colour intensity and the structure of the scene, whereas the depth map represents the Z-distance per-pixel between the camera and a 3D point in the visual scene. The depth map have two main characteristics: has large smooth regions, and has abrupt changes in the objects' boundaries. Normally, the depth is quantised with 8 bits, *i.e.*, allowing 256 different depth levels, with the closest point associated with the value 255 and the most distant point associated with the value 0. Therefore, the depth map is defined as a smoothed grey level representation. This data format was proposed by the European Information Society Technologies (IST) in the "Advanced Three-Dimensional Television System Technologies" project [32].

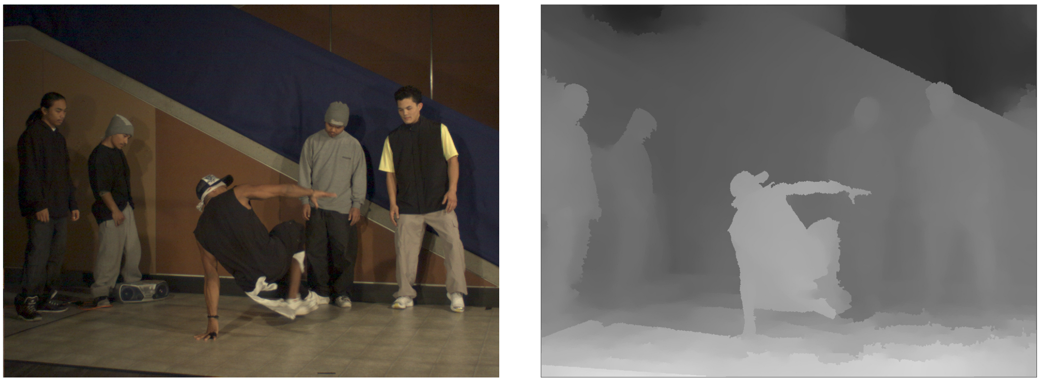


Figure 2.4: Representation of the video plus depth information.

The main advantage of depth-based representations lies in its structure, which allows to generate an arbitrary number of views at the receiver. Therefore, any stereo or multi-view display can be used by such 3D video decoders, as the required number of views can be individually synthesised for each type of display [31]. The combination of both texture and depth information is done using a depth-image-based rendering (DIBR) technique [33,34] to generate synthesised views. Through this technique, the receiver can render virtual views of the 3D scenario. When both views have the correct disparity, the pair can be displayed to provide a 3D viewing sensation to the observers.

The texture and corresponding depth can be encoded separately using a state-of-the-art 2D video encoder, *e.g.*, the H.264/AVC [17]. The H.264/AVC supports the encoding of an auxiliary monochrome picture together with the primary one [24], therefore, both signals can be encoded together, but without exploiting their inherent dependencies. The coding efficiency of this method can be improved by sharing the motion information from the main texture picture. This improves the coding gains and reduces the coding complexity, achieving a reduction of the encoding time up to 60% comparing with the original scheme [35]. Another approach is to encode the depth map based on geometric representation of the texture image [36]. The image is subdivided using a quadtree decomposition and an appropriate modelling function is selected for each region of the image, so, the overall rate-distortion cost is optimised. The benefit of this approach was shown to improve the rendering quality in comparison with JPEG 2000 and also with the H.264/AVC INTRA coding [37]. Some artefacts (*e.g.*, occlusions) may arise on rendering secondary views, thus, sending additional information might be an useful solution to eliminate these artefacts. This additional information is called shapes masks [38]. This coding scenario is fully adapted to 2D television because all streams are coded by 2D video codecs. Backward compatibility is ensured, because older systems just need to discard the depth map information. Switching from 2D to 3D television, using a V+D format, has a bitrate increase of about 10% to 20% [39].

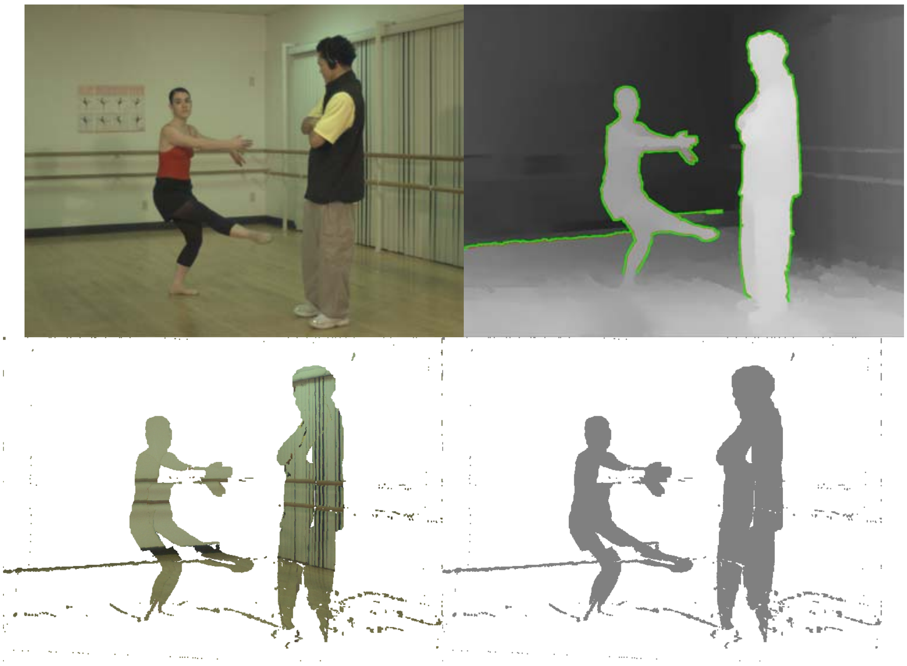


Figure 2.5: Layered depth video representation [20].

To improve the users' experience with a depth-based representation, a 3D system can handle multiple pairs of a 2D image and associated depth map. Increasing the number of pairs, the viewers can see the scenario from different angles. This is the multi-view plus depth (MVD) format, and compared with the MVC extension, MVD does not need to send the same number of views to achieve the same viewing experience [17]. 3D applications based on the MVD format have more 3D viewing positions, which increases the level of flexibility to the display technology. Moreover, since there exist multiple 3D viewpoints motion parallax is possible, extending the level of interactivity.

Another novel representation of the 3D scenes is the layered depth video (LDV) [40]. The LDV is a derivative and an alternative to MVD representation, and can be acquired by warping n depth images into a common camera view. Similarly to the V+D format, this format uses one colour video with associated depth map (*i.e.*, main layer). However it includes a further background layer with the associated depth map. The background layer includes image content which is covered by foreground objects in the main layer. Figure 2.5 illustrates the main layer on the top and the background layer with the associated depth map at the bottom of the figure. Extensions to the LDV may use one or more residual (background) layers, which include data from other viewing directions, not covered by the main view. LDV supports rendering of virtual views and, therefore, multi-view autostereoscopic displays.

LDV is more efficient than MVD with lower coding complexity, but some displaying artefacts may be visible, due to inaccurate conversion between multi-view and LDV [17].

Moreover, LDV format is much more error prone, due to the relevance of the background layer in the view synthesis process. The main drawback of depth-enhanced formats is the very high computational complexity associated with the rendering process. Moreover, there are still some open problems related with depth capture and view synthesis, which need adequate solutions.

The standardisation of the V+D coding is referred to as MPEG-C Part 3 [6] (also referred to as ISO/IEC 23002-3). This specification is based on encoding 3D content into a conventional MPEG-2 transport stream, which includes the texture video, the associated depth maps and some auxiliary data [41]. This standard meets the requirements of the broadcast infrastructure, by providing content interoperability, as well as display and capture technology independence [42]. Since the standard only defines high-level syntax that allows the receiver to correctly decode two incoming video streams, it is flexible to the differen coding algorithms. Another key advantage is that the MPEG-2 bitstream provides backward compatibility with existing 2D systems.

2.2 MVC extension of the H.264/AVC

Compression of the 3D video formats discussed in the previous section is required, to obtain an efficient coded representation. As mentioned before, the state-of-the-art 3D compression technique is the most recent major extension of the H.264/MPEG-4 AVC standard [24] is the MVC extension [5].

In the case of a stereo system, a multi-view profile (MVP) was defined in MPEG-2 standard [17, 43], which allows the transmission of two video signals for stereoscopic TV applications. One of the main features of MVP is the use of scalable coding tools to guarantee backward compatibility with the MPEG-2 Main Profile. The MVP relies on a multi-layer representation such that one view is designed as the base layer and the other view is assigned as the enhancement layer. In the base layer, only temporal predictions are used, while temporal and inter-view predictions are simultaneously exploited on the enhancement layer. As for the MVP profile of the MPEG-2 standard, the MVC extension improves the coding efficiency of multi-view video, by exploiting the redundancy over time and across views. Therefore, pictures are not only predicted from temporal neighbours, but also from spatial neighbours in adjacent views.

In the MVC extension [29] different techniques were introduced, which are able to improve coding efficiency and reduce decoding complexity. In this section several key features of the MVC extension will be described.

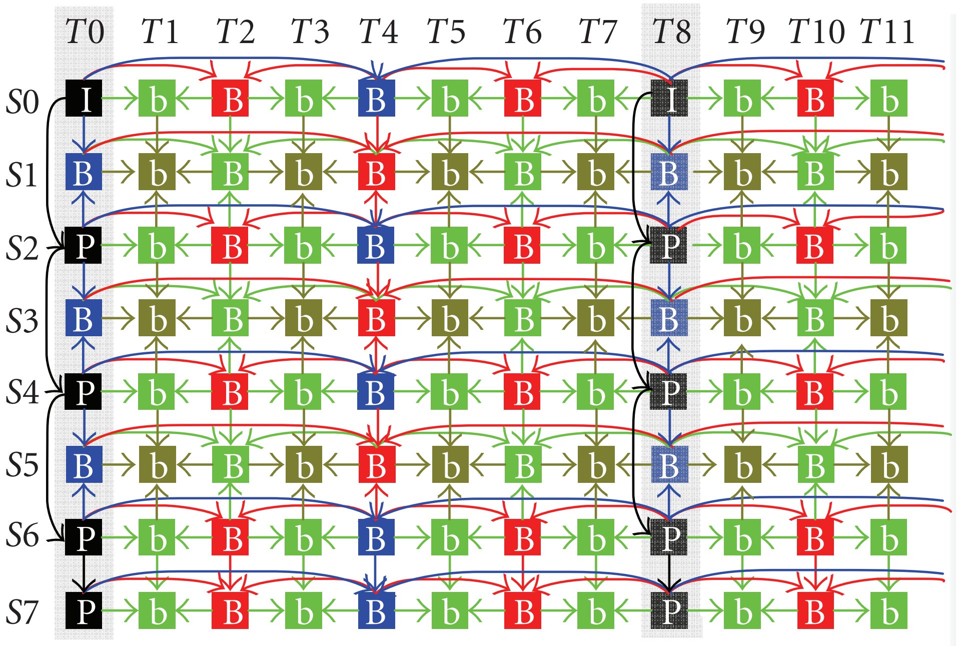

T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11
S0 I b B b B b B b I b B b
S1 B b B b B b B b B b B b
S2 P b B b B b B b P b B b
S3 B b B b B b B b B b B b
S4 P b B b B b B b P b B b
S5 B b B b B b B b B b B b
S6 P b B b B b B b P b B b
S7 P b B b B b B b P b B b

Figure 2.6: Illustration of the MVC prediction structure [44].

View dependencies

A common prediction structure of MVC, using both hierarchical temporal prediction and inter-view prediction, is illustrated in the Figure 2.6. In this scheme, some views depends on other to be accessed and decoded. The example of the figure shows that view 5 depends on both view 4 and view 6. Then, view 4 depends on view 2, and view 2 depends on the base view (*i.e.*, view 0).

Dependencies are defined for each view, and transmitted through the sequence parameter sets (SPS) extension. There exist two types of views in the MVC extension:

- The **base view**, corresponding to the first view of Figure 2.6, is independently coded, as in a simulcast coding, exploiting only temporal redundancies. The base view enables synchronisation and random access features, when the key picture in INTRA coded.

- The **non-base views**, also known as auxiliary views, correspond to the remaining views. In this view both temporal and inter-view dependencies are exploited, by using motion and disparity compensation respectively. Consequently, some coding gains are achieved, but the random access for each view is lost. In order to overcome this problem, *anchor pictures* are introduced. One *anchor picture* does not have any temporal dependency on other pictures from the same view, so it may only be predicted from pictures from neighbour views at the same instant.

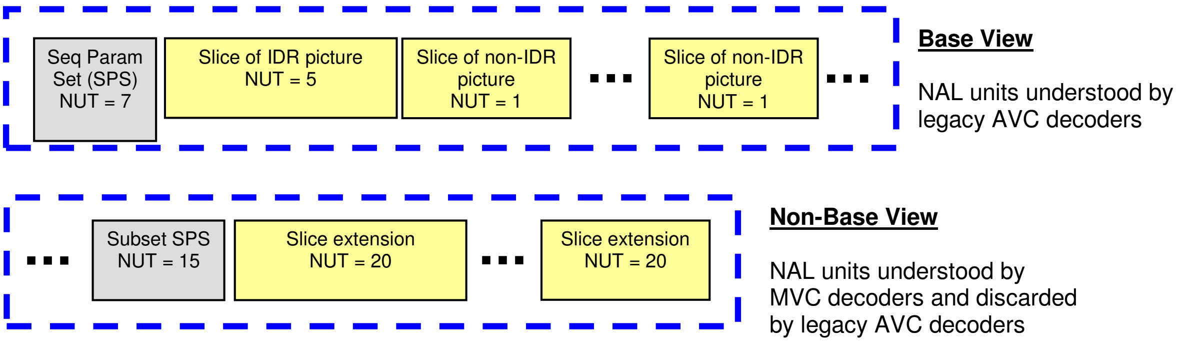

Seq Param Set (SPS) NUT = 7
Slice of IDR picture NUT = 5
Slice of non-IDR picture NUT = 1
Slice of non-IDR picture NUT = 1
Base View
NAL units understood by legacy AVC decoders
Subset SPS NUT = 15
Slice extension NUT = 20
Slice extension NUT = 20
Non-Base View
NAL units understood by MVC decoders and discarded by legacy AVC decoders

Figure 2.7: Structure of an MVC bitstream including NAL units that are associated with a base view and NAL units that are associated with additional non-base views [19].

Bitstream structure

In MVC, it is mandatory for the compressed multi-view bitstream to include a base layer stream that may be independently extracted and used for backward compatibility with legacy 2D devices. The AVC standard defines a network abstraction layer (NAL) where the coded data is organised into NAL units. Different types of NAL units are defined in the standard, some of them used for coded video, while others for high level syntax information. To achieve backward compatibility, video data of the additional views (*i.e.*, non-base views) are encapsulated into an extension NAL unit type that is used for both scalable video coding (SVC) [45] and multi-view video. Figure 2.7 illustrates some of the existent MAL units for the base and non-base views. As shown in the figure, some new NAL unit types (NUT) are introduce to carry the non-base view information, *i.e.*, video (NUT=20) and specific high-level information (NUT=15). The decoding process for the new NAL units is only defined in the MVC profiles. Therefore, the decoders conforming only with single-view profiles will discard the NAL units conforming to a given MVC profile, since it does not recognise them.

Inter-view prediction

The view dependency structure, presented in Figure 2.6, shows an efficient way to improve the compression efficiency. The video data is compressed by exploiting both temporal and inter-view redundancies. In other words, the MVC standard is able to reduce the redundancies among the pictures of each view, and the inter-view redundancies, among pictures of different views. Consequently, additional coding gains compared to the H.264/AVC simulcast approach are achieved [30]. The inter-view correlation between adjacent viewpoints is removed using disparity estimation and compensation. The MVC also enables the use of variable block based motion estimation, as previously used for temporal predictions in H.264/AVC.

According to the MVC specification, inter-view reference pictures must be contained within the same access unit as the current picture, *i.e.*, pictures belonging to the same capture or display instant as the current picture. The MVC design does not allow the prediction of a picture in one view at a given time using a picture from another view at a different time. Thus, the inter-view prediction does not involve different access units, which restricts computational complexity without losing significant coding gains.

Further improvements are also possible, since no geometrical constraints between views have been considered [46]. Moreover, the difference of viewpoints inside a multi-view video sequence may create illumination changes between the views. To compensate the inter-view illumination changes, an illumination change-adaptive motion compensation has been proposed and implemented in [47].

Random access and view switching

Since MVC introduces dependencies between views, random access must also be considered in the view dimension. Specifically, to access a given view (referred to as target view), any view on which the target view depends needs to be accessed and decoded for the purpose of inter-view referencing. Thus, additional processing delay is introduced, when accessing a non-base view.

Instantaneous decoding refresh (IDR) pictures are typical random access points. Any picture succeeding the IDR picture in decoding order (*i.e.*, bitstream order) cannot be inter predicted from earlier pictures within the same view. In the context of MVC, an IDR picture within a given view disallow, at that particular instant, the use of temporal prediction for any view on which a given view depends. Therefore, only inter-view prediction may be used for encoding such pictures. Moreover, in MVC extension an *anchor picture* for a view are also defined. These anchor pictures are identical to IDR pictures, however they allow the use of inter prediction in pictures on which the anchor picture depends. As both IDR and anchor pictures restrict the inter prediction possibilities, the coding efficiency in such pictures decreases.

View switching refers to the change of the target view. The number of target view(s) may be one or more. In case the number of target views changes or any of the target view is changed from one view to another, a view switching occurs [44]. This must happen at view switching points, after which the new target view can be correctly decoded. Free viewpoint video is a typical application where view switching is useful. All random access points can also be used as view switching points. Besides that, more switching points may be added by specifically setting the interview prediction relationship or by using the SP- or SI-slices [48].

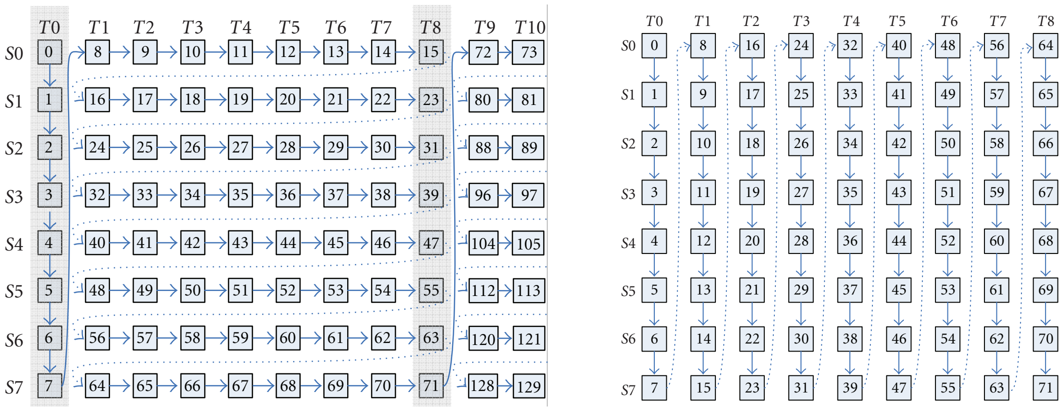

T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10
S0 0 8 9 10 11 12 13 14 15 72 73
S1 1 16 17 18 19 20 21 22 23 80 81
S2 2 24 25 26 27 28 29 30 31 88 89
S3 3 32 33 34 35 36 37 38 39 96 97
S4 4 40 41 42 43 44 45 46 47 104 105
S5 5 48 49 50 51 52 53 54 55 112 113
S6 6 56 57 58 59 60 61 62 63 120 121
S7 7 64 65 66 67 68 69 70 71 128 129
T0 T1 T2 T3 T4 T5 T6 T7 T8
S0 0 8 16 24 32 40 48 56 64
S1 1 9 17 25 33 41 49 57 65
S2 2 10 18 26 34 42 50 58 66
S3 3 11 19 27 35 43 51 59 67
S4 4 12 20 28 36 44 52 60 68
S5 5 13 21 29 37 45 53 61 69
S6 6 14 22 30 38 46 54 62 70
S7 7 15 23 31 39 47 55 63 71

Figure 2.8: View-first coding (a) and time-first coding (b) approaches [44].

Decoded picture buffer management

The MVC extension makes an efficient use of the multiple reference frames feature of H.264/AVC, and more specifically of the decoded picture buffer (DPB), by inserting and deleting frames from adjacent views among the temporal reference frames. In H.264/AVC, the order by which NAL units are placed inside the bitstream is referred to as the decoding order. In multi-view video, where two dimensions, time and view, are involved, the decoding order concept becomes more complex. Two different decoding order arrangements can be considered: either view-first coding or time-first coding, as shown in Figure 2.8. In view-first coding, coded pictures belonging to different views but with the same time instance are interleaved with pictures from other time instants. Thus, they cannot be in the same access unit. Consequently, view-first coding requires a processing delay proportional to the group of pictures (GOP) size multiplied by the number of views. Furthermore, the possibility of parallel decoding [44] is compromised in this coding option.

Time-first coding in the MVC extension is used, in order to overcome the previous problems [49]. In time-first coding, pictures of the same temporal instants are contiguous in decoding order. Therefore, one access unit contains pictures from different views belonging to a given instant. For example, if the delay between the picture decoding and output is not taken into consideration, in the H.264/AVC using hierarchical B pictures the optimal DPB size is $TL+1$, where TL is the highest temporal level of the hierarchical B coding pictures and is given by, $TL = log_2(GOP\ length)$ [50]. In H.264/MVC, for the typical prediction structure illustrated in Figure 2.6 the maximum DPB required to encode nv views in time-first coding is $nv \times (TL + 1) + 2$. This is less than the size needed for the view-first coding [51]. Moreover, when the output is considered and the picture should be stored in the buffer until the display instant, the view-first coding case is even worse, especially for multi-view video applications when it is required display all the views.

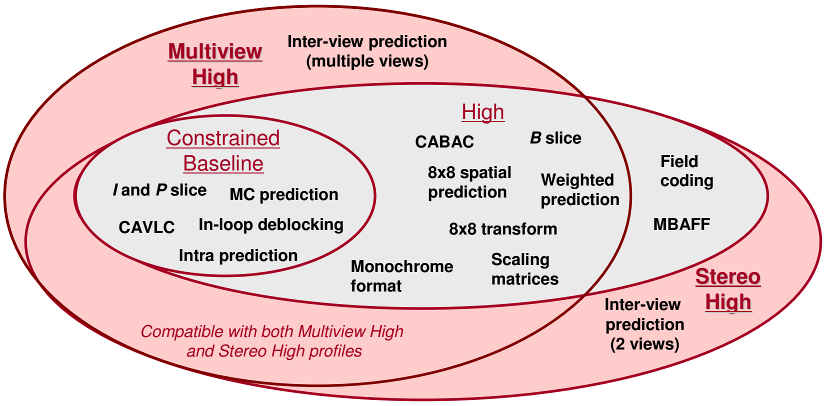


Figure 2.9: Illustration of MVC profiles, *i.e.*, Multiview and Stereo High profiles, and the relation with the High and the Constrained Baseline profile [5].

MVC profiles

The MVC extension defines two new profiles to determine the subset of coding tools that must be supported by compliant decoders. These tools are defined in order to cope with the requirements of the 3D/multi-view video. The profiles for the MVC support one or more views and are both based on high profile of the H.264/AVC, with a few modifications [5]. The Multiview High profile supports multiple views but does not supports interlace coding tools; the Stereo High profile is limited to two views, but support interlace coding tools.

In both profiles, the base view can be encoded using either the High profile or the recently added Constrained Baseline profile [28] of the H.264/AVC. An illustration of the two novel profiles introduced with the MVC extension is shown in the Figure 2.9. The figure also shows the previous profiles of the H.264/AVC. Note that, if the High profile is used in the base view, the interlace tools (field picture coding or macroblock-adaptive frame-field - MBAFF), which are commonly supported in the high profile, cannot be used in the base layer of the MVC stream, since they are not supported in the Multiview High profile.

2.3 Transport and transmission of 3D video contents

Three-dimensional video is currently being introduced to the home through various channels, including broadcast via cable, terrestrial and satellite transmission, streaming and download over the Internet, as well as on storage media such as Blu-ray discs. Figure 2.10 shows different system technologies, *e.g.*, MPEG-2 transport stream, RTP or ISO file format, each one associated with one of the delivery systems mentioned before. The system for delivery of 3D video is an important component in the 3D video systems data

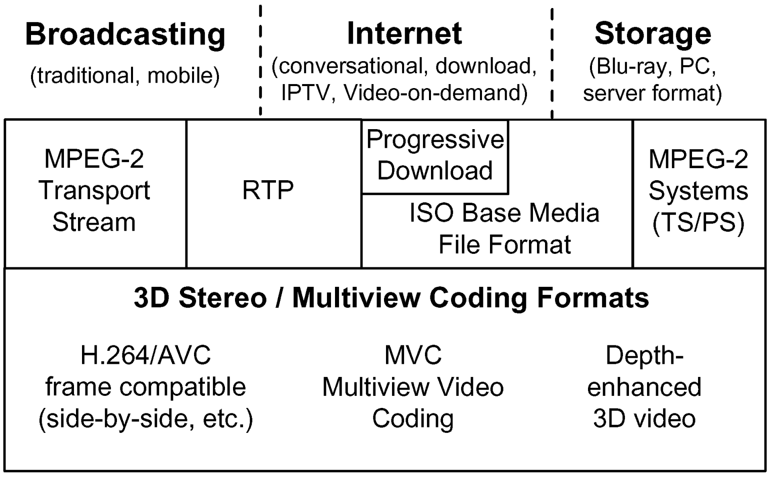

Broadcasting
(traditional, mobile)
Internet
(conversational, download, IPTV, Video-on-demand)
Storage
(Blu-ray, PC, server format)
MPEG-2 Transport Stream
RTP
Progressive Download
ISO Base Media File Format
MPEG-2 Systems (TS/PS)
3D Stereo / Multiview Coding Formats
H.264/AVC frame compatible (side-by-side, etc.)
MVC Multiview Video Coding
Depth-enhanced 3D video

Figure 2.10: High level overview of 3-D transport and storage systems [52].

processing chain and allows for its successful deployment into various electronic devices. In this section different approaches and techniques used to transport the 3D video are introduced.

Different approaches to represent and code 3D video were presented in the Section 2.1, and, as shown in the Figure 2.10, they all can be used in the transport and storage systems. The MPEG-2 System [53] is used globally for digital television broadcast and optical disc storage. The standard defines the multiplexing and synchronisation of compressed video, audio, and auxiliary information [54]. In the MPEG-2 transport stream, the data stream is encapsulated into fixed length transport packets, which provides more error resilience for broadcast scenarios such as terrestrial and satellite systems. Moreover, it is capable of carrying different formats used for 3D multimedia.

The recent amendment of the MPEG-2 systems standard for MVC defines the transport of one or more views in different streams [55], referred to as MVC sub-bitstreams. The combination of such sub-bitstreams results in a suitable MVC stream as defined in [29], restricting only a single view, or a subset of subsequent views in decoding order, as part of MVC sub-bitstreams. This keeps the concatenation process simple at the receiver point. Moreover, as referred in [52] there are three different ways of transporting the base and non-base views. The first one consists in transport all views within the same packetized elementary stream (PES) and with the same program ID. The second is to transporting just one view in one PES of the same program. In the third method the base view with one or more non-base views are transported in separate PES within the same program. The high flexibility of the MPEG-2 TS makes it the suitable container format for the transport and storage of 3D contents, as described in [56]. Moreover, as it allows different multiplexing approaches for 3D video, it supports unequal error protection and hierarchical transmission channels.

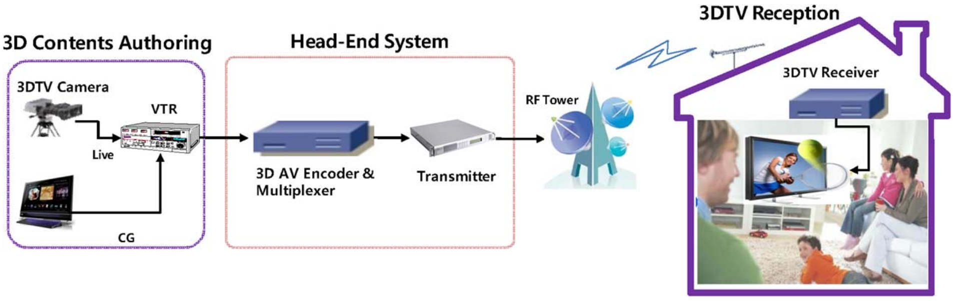


Figure 2.11: Illustration of a terrestrial 3DTV broadcasting system [15].

In terms of storage of the 3D video the Blu-ray Disc fulfils the requirements. Nevertheless, in order to guarantee the video quality, a frame sequential full-resolution stereo video format was first considered. However, subjective evaluation results leaded to the inclusion of the MVC stereo profile as the mandatory 3D video codec in the Blu-ray 3D specification [57]. The backward compatibility with legacy 2D players is maintained using two MPEG-2 TS: a main TS for the base-view and associated audio, and a auxiliary sub TS for the non-base view and other elementary data associated with the 3D display, such as the depth of the subtitles. This approach is able to achieve a maximum video bitrate in the stereo video of 60 $Mbps$ and a maximum per view bitrate of 40 $Mbps$ [18]. Nevertheless, in Blu-ray 3D standard, both single and dual TS are defined. The single TS is used when a 3D bonus video is encoded at a lower bitrate, or when an 2D video clip is encoded using H.264/MVC to avoid switching between AVC and MVC decoding [18].

The development 3D video broadcasting systems has several challenging issues: sharing the transmission bitrate between views, choosing the approach to multiplex the video data in order to maintain the backward compatibility with the available equipments and adaptation with the receiver end (*e.g.*, set-top boxes and displaying technologies). In Figure 2.11 the general concept of a terrestrial 3DTV broadcasting system is represented. In [15] an experimental 3DTV service was accomplished, following the overview diagram of the Figure 2.11, coding the base view using the traditional MPEG-2 encoder and non-base view with the H.264/AVC. This approach maintains the compatibility with the MPEG-2 decoders, while achieving more efficiency in the non-base view coding. Although the 3D service managed to be delivered over the available bandwidth, reducing the bitrate of the 2D video (*i.e.*, base view), did not exploit the spatial redundancy between the left and the right view.

A more sophisticated delivery of MVC video over MPEG-2 TS is presented in [56]. Based on the target view and the prediction structure used at the encoder it is proposed a priority-based selective transport framework. Through the priority-based system view scalability is achieved effectively, and it is possible to satisfy different requirements of the multiple views under different application scenarios of multi-view video.

2.4 Summary

This chapter presented an overview of the main 3D video formats used for broadcast. The advantages and drawbacks of each format were mentioned, as well as some coding techniques used for each format. Although stereoscopic video is one of the oldest 3D technologies, it is still a very popular format used in many proposals recently presented to code and transmit 3D video. The depth-based representations are efficient alternative formats, especially for the multi-view applications, since they can significantly reduce the amount of data for the same quality of experience. However, the main goal of the research work is to investigate the problems of error concealment in the 3DTV services, and the stereoscopic video format seems to be a suitable format for this kind of application. Chapter 4 will present a subjective assessment study to evaluate the impact of temporal distortion caused by missing frames in the stereo video at the receive side. Moreover, Chapter 6 also focuses in the stereo video, describing different concealment schemes to cope with whole frame loss.

Chapter 3

Error concealment: a review

Digital transmission through networks is, in general, prone to errors and data loss. These errors can be handled in different ways, such as retransmission, error correction, and error concealment [58, 59]. When an encoded video bitstream is erroneously received in the decoder side, *i.e.*, when error protection or error correction methods are not able to avoid all errors, the erroneous artefacts in the video are mitigated via methods referred to as error concealment. In general, the error concealment operation is a post processing step in the decoder side and it is the last step to handle errors in the communication chain. Thus, concealment of missing data should take place within the video decoder to minimise perceived effects of lost data. In summary, to enhance the 3D video quality and depth perception, the design of the transmission system must involve the development of adequate error concealment tools [60].

This chapter reviews several error concealment techniques used to enhance the quality of experience both in monoscopic and stereoscopic video, being this last the main subject of this thesis. The monoscopic (*i.e.*, 2D video) loss concealment algorithms presented in literature are worth to be mentioned, as they provide relevant insights, and possible exploitations of information in spatial and temporal domains, that may be useful in the proposal of a novel error concealment algorithm for 3D video. Then, concealment strategies to recover missing data in 3D video are also detailed. Section 3.1 reviews several concealment methods based on spatial redundancies within the lost frame, describing some algorithms based on pixel interpolation, and others based on frequency domain information. In Section 3.2 temporal error concealment techniques are described, including the well-known boundary match algorithm and the motion vector extrapolation. Section 3.3 introduces some concealment methods applied to stereoscopic video, extending beyond the traditional ones for 2D video by using the inter-view dependencies. Finally, Section 3.4 concludes this chapter.

3.1 Spatial error concealment techniques

This section introduces several concealment techniques that exploit the spatial redundancy of images to recover the lost data. To recover lost information, the methods introduced in this section entirely rely on neighbouring image regions received without errors. Consequently, spatial error concealment is mostly useful in the case where information of the neighbouring frames are not available.

3.1.1 Pixel level concealment

A typical approach to conceal missing regions in an image is based on linear interpolation of neighbouring pixels [61,62]. Figure 3.1 shows the estimation process of a lost pixel (lp), within the lost region (R_L), using the four closest pixels (p_i) as reference, and the distance between the lost pixel and the reference ones (d_i). The pixel value (lp) is recovered using a weighted average, as follows:

$$lp = \sum_{i}^{n} \frac{p_i}{d_i} \Bigg/ \sum_{i} \frac{1}{d_i} \,. \tag{3.1}$$

Liner interpolation is a simple yet effective method to recover lost regions in digital images. However, as each lost pixel is recovered by giving the same importance to each reference, the image discontinuities are not taken into account. Thus, some pixels used as reference may belong to distinct regions of the image, without significant correlation with the lost region itself. This leads to a poor estimation of lost pixels, resulting in noticeable artefacts in the reconstructed image. For example, in Figure 3.1 the lost pixel (lp) and the pixel p_4 are separated by an edge. Therefore, p_4 may not be suitable to recover the lost pixel value. So, to ensure a correct concealment of the lost region, the edges and image transitions must be considered, by selecting correct pixels to be used in the interpolation function. To overcome this problem, a directional interpolation can be performed, by recovering the lost pixels based on those within the same region, according to the edge directions.

The directional interpolation improves the performance of the pixel recovery, by estimating the lost edges using either correctly received or previously recovered pixels. In order to obtain the best direction to interpolate, edges are calculated using the image gradient, for example applying the Sobel operator [63]. This operator gives the first-order approximation of the gradient of the image. This is widely used in image processing to

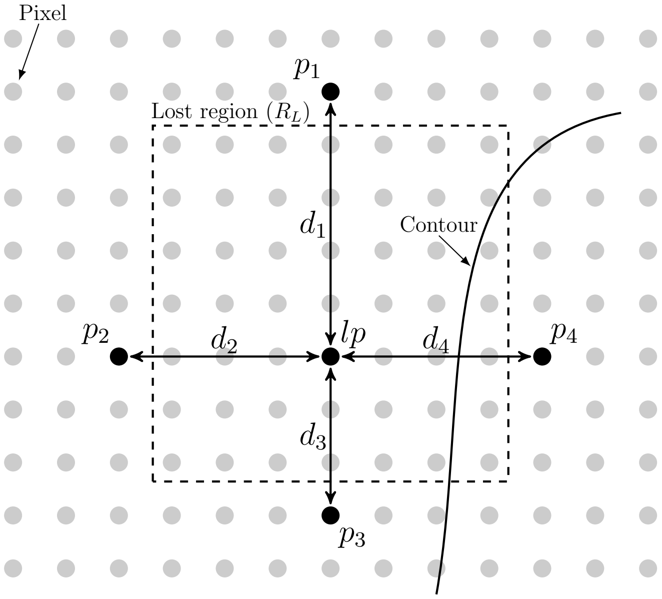


Figure 3.1: Lost pixel interpolation from neighbouring error-free pixels.

locate the image edges, and uses an horizontal and vertical mask given by:

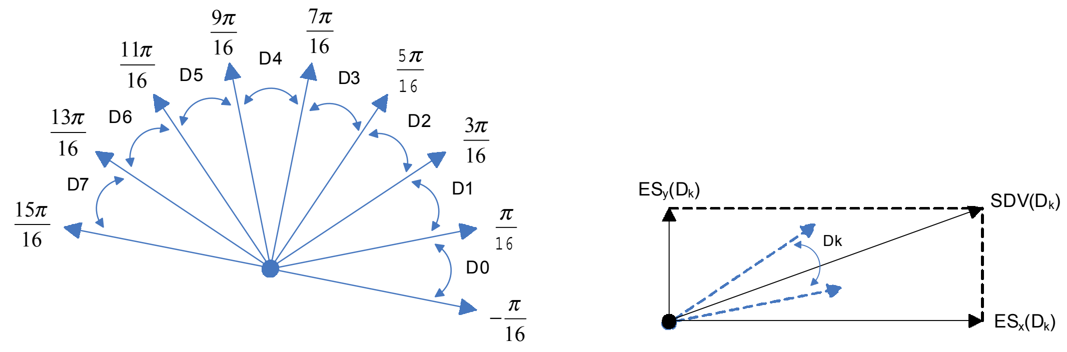
$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{3.2}$$

The convolution between the pixels and the Sobel matrixes results in both horizontal (g_x) and vertical (g_y) gradients, from which the magnitude and the angular direction are computed using,

$$\begin{aligned} G &= \sqrt{g_x^2 + g_y^2}, \\ \theta &= tan^{-1}\left(\frac{g_y}{g_x}\right). \end{aligned} \tag{3.3}$$

The Sobel operator is an efficient algorithm to detect edges, as it can reach extreme values at the image edges, so it can be used in pixel interpolation. The usage of this operator for error concealment was evaluated in [64–68], and it is characterised as a fast and efficient method due to its small kernel [69].

In order to extract the best direction for interpolation, the Sobel operator is used in a region around the missing pixel. In [64] the gradient magnitude (G) is compared against a threshold $\alpha = 90$, and only directions with a gradient higher than α are considered for analysis. The direction (θ) of the chosen candidates are quantised in steps of 10°, and an edge direction is marked as dominant if its frequency of occurrence is higher than the other ones. A dominant direction is detected if the difference between the two largest frequencies is greater than 10% of the largest frequency value.



(a) (b)

Figure 3.2: (a) Eight possible edge directions to recover the missing block; (b) definition of spatial directional vector based on edge strength [65].

The procedure described above can be improved with refinement, recovering a more reliable edge information, which leads to higher quality in interpolated pixels. In [65], after detecting the direction of edges using the Sobel masks, a quantisation is performed using 8 directions, as present in the Figure 3.2(a), so the analysis can be focused in those possible directions. In order to extract the edge strength, the sum of the g_x and g_y (*i.e.*, result of the Sobel operator) for each direction shown in the figure is calculated. Figure 3.2(b) shows the definition of the spatial directional vector (SDV) for the direction D_k, based on both horizontal and vertical edge strength (ES). Although the SDVs are used to discover the optimal edge direction, only the largest vectors represent a relevant edge, thus, a dynamic threshold is defined by $T_{ES} = 0,5 \times max \{ES(D_k)\}$, in order to discard the SDVs lower than T_{ES}. Although the concealment method presented in [64] is not very sensitive to the threshold value applied to the gradient magnitude, a dynamic threshold is expected to perform better, since it can adapt to each image. The set of remaining SDVs are evaluated to find the best directions. This is done by calculating the difference between the two closest pixels within the line that passes through the lost pixel and has the direction of the corresponding SDV. To ensure the fidelity of the concealment, the distance between the two pixels is restricted within twice the block width. The differences between pixels are analysed and the two SDVs with the small directional distortion are selected to reconstruct the missing block. Then, the interpolation uses the following iterations:

1. Determination of the contribution of each SDV, applying the Equation 3.1 to the two closest neighbour pixels along the SDV;

2. Restore the pixel, by averaging the contributions from the two SDVs.

In [65] the previous method is compared with other concealment methods, *e.g.*, the method presented in [61], in different loss conditions. The results showed a good perfor-

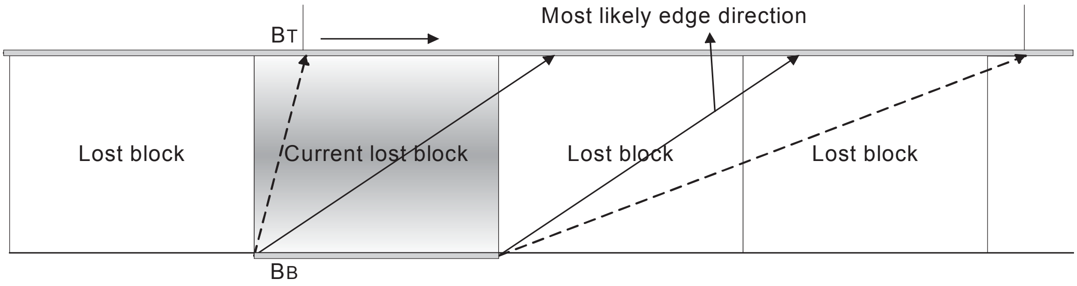


Figure 3.3: Boundary matching algorithm used in the modified edge-oriented concealment method presented in [67].

mance comparing with linear interpolation techniques, showing the relevance of directional interpolation. Moreover, the previous work tested the concealment method using a single SDV, and a weighted average of two SDV contributions. Although the weighted factors are related with the directional distortion, the simple average proved to be more efficiency.

In [66] a method to conceal consecutive missing blocks based on one-dimensional (1D) boundary matching is presented. Figure 3.3 shows an example of consecutive lost blocks. The boundary matching is used to connect the top and bottom pixel information, which are the only information available without errors. The boundary match finds the best correlation between the top and bottom boundary, by minimising the error, *i.e.*, the absolute difference (AD), between the top and bottom pixels measured as follows:

$$AD(x) = \sum_{i=0}^{N-1} \left| f_i^{B_T} - f_{i+x}^{B_B} \right|, \tag{3.4}$$

where the $f_0^{B_T}$ and $f_0^{B_B}$ is respectively the first pixel of the top and bottom boundary of the current lost block. The N is the size of the block and x represents the shift from the top to the bottom boundary. Equation 3.5 is also applied to the bottom boundary, which results in two possible edge directions. Those directions (see the diagonal lines in Figure 3.3) are used to recover the missing pixels within the lost block, using a directional weighted interpolation along previous determined directions. For the pixels with overlapped edges, the average of the two linear interpolations is used.

The performance of the method presented in [66] can be improved by determining the most likely edges using the Sobel operator. This operator is widely used in spatial concealment methods as showed before in this section, achieving acceptable quality in the reconstruction of the missing pixels [65]. In [67] the combination of the Sobel edge detection with the boundary matching algorithm was proposed. The first step is the determination of the most likely edge direction (see Figure 3.3). Then, the boundary match is performed around the most likely direction, decreasing the probability of miss-detection of the edge direction. Although the cost function to determine the final direction is based on the Equation 3.4, it takes into consideration the distance between the new candidate (x) and the most likely edge direction (x_0), as shown in following equation:

$$COST(x) = AD(x) + (N/2) \times |x - x_0| \,. \tag{3.5}$$

A comparison with the method proposed in [66] is performed. Results show that method [67] is able to efficiently recover most significant edges, achieving higher subjective quality and an average objective gain of 1 dB in PSNR.

The H.264/AVC coding standard [24] exploits the redundancy of the INTRA coded frames using a spatial prediction. The standard defines different INTRA prediction modes with three different block sizes: 16×16, 8×8 and 4×4, and different directional modes. In [70] the amount of INTRA predictions that match the image's edges was measured, revealing that there is a strong correlation between the edges and the chosen methods in the encoder. Thus, spatial interpolation for error concealment can be based on the coding modes (*i.e.*, direction of INTRA prediction), instead of relying on the Sobel operator. This reduces the computational load, maintaining an acceptable number of possible directions, *i.e.*, 8 different directions in case of H.264/AVC coding standard.

The method presented in [65] uses the average of two contributions provided by two edge directions. However, when two relevant edge directions are present the optimal approach is to combine the detected directions in order to select a more appropriate group of pixels to interpolate the missing ones. In [70] a method to combine different edge directions is assessed, selecting the dominant regions to perform a sequential directional interpolation as described below. Since this method uses the H.264/AVC coding modes (presented above) to detect the edges, it firstly needs to connect the correlated blocks. This is achieved by comparing the edge attributes, *i.e.*, horizontal and vertical magnitude and angle. The blocks with the highest similarity (smallest difference between attributes) are connected and so on, until all blocks are connected, creating an edge direction. Figure 3.4(a) illustrates an example, where the main edge region is represented with dark grey. The unmatched blocks are connected, by expanding the line within the edge direction into the lost MB until reach a previously connect edge. This edge region is also represented in the figure with light grey. Figure 3.4(a-f) represents each step of the sequential interpolation method proposed in [70]. The lost pixels within the MB are interpolated with the following order: (i) the pixels within the edge areas are interpolated using the edge direction; (ii) the pixels in the flat areas are recovered using the closest pixels within the same region. Firstly, the pixels in dominant edge direction, represented with dark grey in Figure 3.4(b), are interpolated using the closest pixels within the edge direction represent with black squares in the figure. Since the subsequent edge direction meet the dominant edge, to perform a correct interpolation the pixel p_2, three pixels belonging to

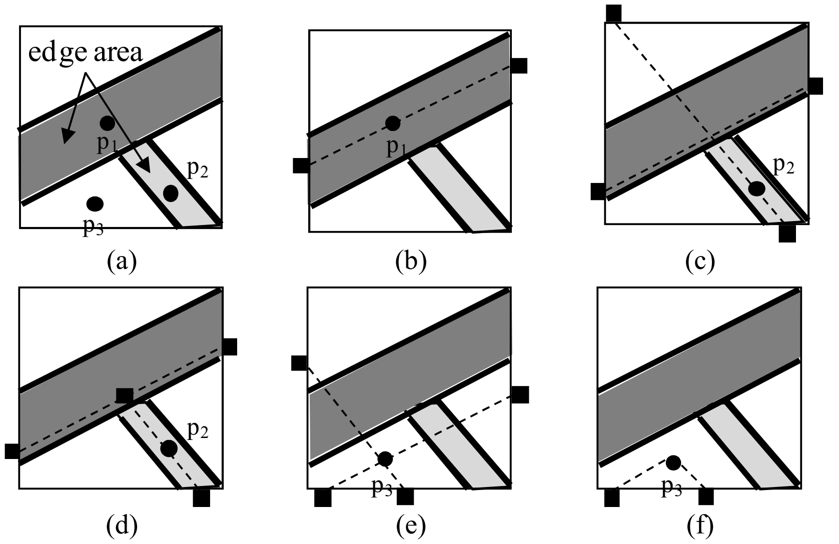


Figure 3.4: Example of detected edges (a); sequential directional interpolation of the lost block (b-f) [70].

the two edges direction are chosen. These pixels are shown in Figure 3.4(d). Finally, the flat areas are interpolated using the surrounding edge directions. An example of this case is the pixel p_3. To interpolate this pixel, instead of using the four pixels illustrated in Figure 3.4(e), only two reference pixels are selected, as represented in Figure 3.4(f). This guarantees a smooth recovery of lost pixel with that region, since the pixels belonging to the opposite side of edges directions are not considered.

This method is an efficient alternative to the Sobel operator, which increases the computational complexity of the concealment method. Moreover, it is able to achieve higher objective and subjective quality than the method presented in [65]. It is also able to significantly reduce the average computation time up to 39% when comparing with the geometric-structure based error concealment proposed in [71]. However, this method is highly dependent of the H.264/AVC coding modes. Therefore, it may fail when consecutive block loss occurs, which is not taken into account in the experimental results presented in [70].

3.1.2 Frequency based spatial error concealment

In this Sub-section the spatial error concealment algorithms that use the dependency between the correctly received information and the missing one in the frequency domain are addressed. These methods reconstruct the missing information maintaining the smooth propriety of the images.

The work presented in [72] considers the reconstruction of a damaged block from the received coefficients and its boundary information. The proposed method makes use of

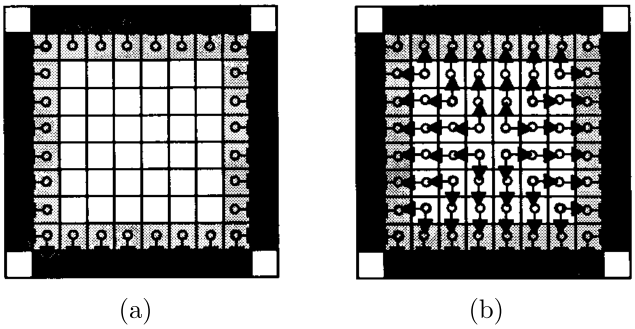


Figure 3.5: Illustration of smoothing constraints [72]: (a) smoothing constraint imposed only on the boundary; (b) smoothing constraint imposed on each sample in direction towards its nearest boundary.

the smoothness property of common image signals and produces a maximally smooth image with the received coefficients and boundary pixels, as in Figure 3.5. This figure shows two possible scenarios. Figure 3.5(a) corresponds to those cases where only the DC coefficient is lost, so the smoothing constraints are only applied to the boundary pixels. Figure 3.5(b) illustrates the smoothing constraints applied when multiple coefficients are lost (a smoothing constraint is applied to the neighbour pixel pointed by arrows). The boundary information is propagated into the damaged blocks such that the transitions along block boundaries are as smooth as possible, so the optimal solution is obtained by two linear transformations. When several spatially adjacent blocks are damaged, initial reconstructions of these blocks are first accomplished by the direct inverse transform with the missing coefficients substituted by zeros. Then, the proposed technique is repeated several times, using the previously reconstructed values as the boundary information in the next step. Simulation results in [72] show that the proposed algorithm is very effective for recovering the DC and low-frequency coefficients. However, when high frequency coefficients are lost, the concealment scheme produces results similar to that achieved by replacing the missing coefficients with zeros, *i.e.*, both will blur sharp edges.

Projection onto convex sets (POCS) method is based on convex sets, which are derived by requiring the recovered block to have a limited bandwidth either isotropically (for a block in a smooth region) or along a particular direction (for a block containing a straight edge). The POCS method is an iterative spatial concealment method [73]. Figure 3.6 represents the block diagram of a conventional POCS algorithm. As shown in the figure, the POCS method is an iterative algorithm, so an initial block ($n = 1$ on the figure) needs to be defined, in order to be combined with eight uncorrupted neighbouring blocks. This forms a 24×24 enclosing block. As shown on top of Figure 3.6, the enclosing block is transformed (DCT) and the high-frequency DCT coefficients outside of an enclosing radius are forced to zero (Pruning). Finally, the enclosing block is inverse transformed (IDCT). The 8×8 block is retrieved, transformed again and the correctly received coefficients are

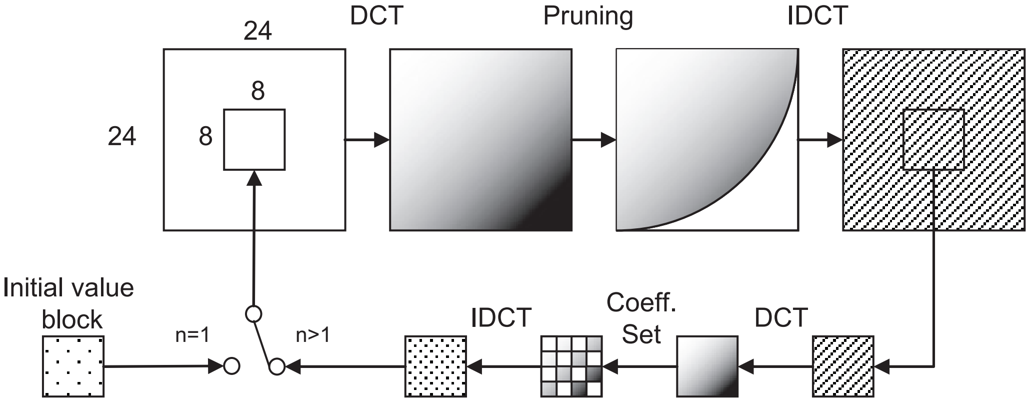


Figure 3.6: Conventional projection onto convex sets block diagram [74].

set in their positions. Finally, after an inverse transform the block is used in the next iteration ($n > 1$ on the figure). This is done until no relevant changes occur between iterations. This method follows two typical proprieties of the video images: smoothness and edge continuity.

The experimental results in [74] show that the performance of the POCS algorithm clearly depends on the relation of the initial block with the missing one. However, only the neighbours blocks are present to be used as reference. In that work the DCT coefficients of the neighbouring blocks are used to fill the missing coefficients and generate the initial block. The sum of the differences between the correct coefficients in the missing block and the received coefficients in the neighbouring blocks is determined, to find the more suitable neighbour to be used as reference. This way, the missing coefficients are replaced with the ones of the neighbour block with the minimum difference, so an inverse transform can be applied to obtain the initial block. This method is compared with the method proposed in [75], which uses the weighted sum of the coefficients of the neighbouring blocks to fill the initial block, achieving higher quality with lower computational complexity [74]. Nevertheless, although the method was tested for the H.263 INTRA, it is noticed that it can also be applied for the integer transform in H.264/AVC standard.

3.2 Temporal error concealment techniques

This section addresses temporal error concealment techniques used to restore corrupted blocks by exploiting temporal correlation between successive frames. Moreover, contrary to the spatial error concealment, the temporal approaches can cope with full frame loss, which may occur in packet video networks. The important issue in these concealment methods is to find the most suitable substitute blocks from the previous frames, *i.e.*,

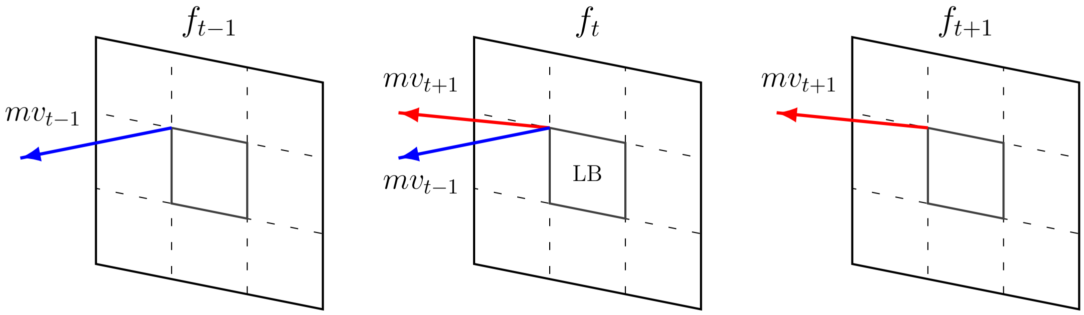


Figure 3.7: Recovery of the frame f_t using the bi-directional motion-copy algorithm proposed in [77].

which may also be posed as a problem of selecting the optimal motion vectors (MVs) for the corrupted blocks.

The simplest way to recover missing frames in the temporal domain is using the Frame-Copy (FC) or the motion-copy (MC) method. These two basic concealment methods were proposed for the H.264 reference software in [76]. The FC method consists in repeating (copying) the temporally closest frame in order to fill the gap of the missing frame. The MC method uses the motion information of the closest frame to recover the missing one, thus the missing pixels are recovered through motion compensation. The experimental results showed that the motion-copy method achieves higher quality in the reconstructed frames in most of the cases. However, the complexity increases with the implementation of the MC method, when compared with the FC method.

In order to improve the motion-copy performance, two reference frames can be used as sources of MVs to reconstruct the missing one. Since the missing frame is only detected after a new frame is received, the missing information can be recovered using the MVs of the previous decoded frame and the subsequent one. Figure 3.7 represents the bi-directional MC algorithm presented in [77]. As shown in the figure, this algorithm recovers the lost MV of a given lost block (LB) in frame f_t using the average value of the MVs available in the frame f_{t-1} and f_{t+1}. The average value of two vectors, mv_{t-1} and mv_{t+1}, is used to reconstruct the missing MV. Then, the missing block is recover using motion compensation. In [77] this method is compared with MC method, achieving gains up to 4.99 dB (PSNR), and up to 0.91 dB over a low complexity pixel-based motion extrapolation method [78]. Moreover, results show that this bi-directional MC method is also able to reduce error propagation.

In [79] a whole frame loss error concealment algorithm was proposed. This algorithm aims to improve the performance of the MC method, since no refinement is performed in the MVs obtained from the closest decoded frame. Moreover, in this work the MVs between consecutive frames are assumed to have a strong correlation, so the motion information of several preceding frames is taken into account to estimate the lost one.

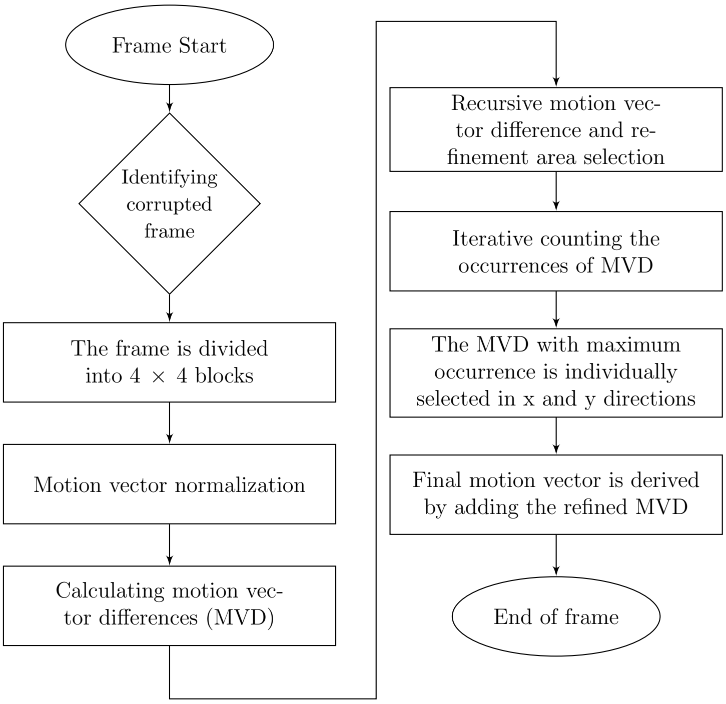

Frame Start
Identifying corrupted frame
The frame is divided into 4 × 4 blocks
Motion vector normalization
Calculating motion vector differences (MVD)
Recursive motion vector difference and refinement area selection
Iterative counting the occurrences of MVD
The MVD with maximum occurrence is individually selected in x and y directions
Final motion vector is derived by adding the refined MVD
End of frame

Figure 3.8: Flowchart of the recursive motion vector refinement concealment method presented in [79].

Figure 3.8 shows the flowchart of the proposed concealment method. As shown in the figure, the erroneous frame is first divided into 4 × 4 blocks and the available motion vectors are normalised, as they may point to different reference frames. Then, the motion vector differences (MVD) between the vectors of the successive frame pairs are recursively calculated, as well as the next refinement area is determined, until it becomes unchanged. Subsequently, the occurrences of a MV difference within each area are counted in x and y directions and the difference with the maximum occurrence in each direction is added to the MV provided by the MC method, thus defining the final refined MV. This is shown to improve the performance of the MC algorithm, achieving gains up 0.47 dB (PSNR) for a packet loss ratio of 20%.

3.2.1 Boundary match algorithms

The recovery of missing blocks in video frames may rely on the available neighbouring macroblocks (MB), to find the suitable MV to reconstruct the missing pixels. A low complexity approach may use the boundary information of surrounding blocks to search for the motion vector based on a spatial smoothness constraint on boundaries of a lost MB [80]. Figure 3.9 represents the boundary information used to estimate the MV for the lost block from a set of different MVs candidates. The boundary match algorithms (BMA) uses the available MVs provided by the top (T), bottom (B), left (L) and right

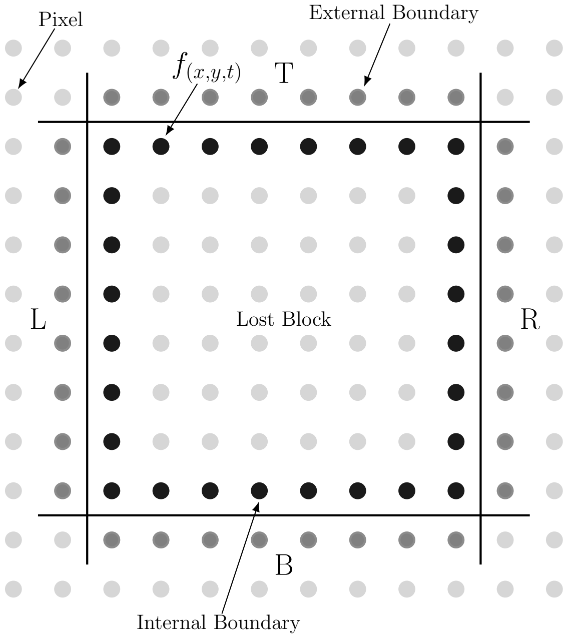


Figure 3.9: Illustration of the boundary match algorithm.

(R) blocks (see Figure 3.9) to fetch the pixels from the previous decoded frame. Then, the optimal MV is selected according to the sum of differences, between pixels on the internal boundary in the recovered block and the corresponding ones on the external boundary. These pixels are marked in the figure.

In [81], the error concealment based on the boundary matching algorithm is improved, by introducing the concept of the interpolated candidate motion vectors (ICMV). The candidate MVs are calculated with an interpolation of the motion vectors of the neighbouring macroblocks in the missing frame and macroblocks around the corresponding block in the previous frame. The ICMVs are created based on the change of magnitude and angle of the top, bottom, left and right neighbouring MVs of the lost block from the previous frame to the current one. Then, these modifications are applied to the MV of the co-located block with the lost one in the previous frame. A more generalised proposal is to apply such transformation not only to the closest neighbours but also to a wider area of blocks. Then, the boundary matching algorithm may choose among these candidates. The proposed method is concluded to be better than the methods with normal candidate motion vectors, especially in high motion video sequences. Moreover, the concept of ICMVs can be applied to any method based on BMA, improving its performance, by introducing more accurate candidates.

Another improvement to BMA was obtained with the Outer Boundary Matching Algorithm (OBMA) [82], also known as Decoder Motion Vector Estimation (DMVE) algorithm in [83]. In this method the distortion is calculated by the difference of two outer boundaries of reconstructed MB, instead of internal and external boundary in BMA. Although

BMA and OBMA have the same design principle of minimising the mismatch distortion, in OBMA the distortion function is measured in the outer boundaries, *i.e.*, external boundaries of the reconstructed block. It was observed in the experimental results that OBMA has better performance in video recovery [82] than conventional BMA. The performance improvements are explained by the usage of the outer boundary in the distortion functions, which incorporates the edge information within the neighbourhood of the lost block in the decision of the most accurate MV to recover the missing pixels [82]. Moreover, in the method presented in [82] further MV candidates were introduced, a refined local search is performed and multiple boundary layers, *i.e.*, increase the overlapped region to measure the distortion and to select the final MV, were tested. The experimental results revealed that the usage of a single layer match gives better performance than multiple layers, while selective search gives excellent trade-off between the image recovery and computation complexity.

In [84] a novel approach to measure the spatial distortion on boundaries of the reconstructed block is presented. Instead of using only horizontal and vertical directions, diagonal and anti-diagonal are also tested, and the best direction is chosen from the reference frame. Since block boundaries may change abruptly, instead of minimising the distortion function the proposed method aims to approximate this distortion to an optimal difference. This optimal difference are determined in the neighbourhood of the co-located block in the reference frame. In [80] a novel distortion function was proposed, to choose among differen candidate MVs. The distortion function exploits both spatial and temporal smooth constraints. The temporal function is based on the outer boundary distortion, as in [82]. The spatial distortion aims to reduce the gradient variances in boundaries of the reconstructed block, so the image edges may be guaranteed. Several experimental results were obtained for different loss events, covering from isolated block lost to the loss of a complete row of blocks. The proposed method is able to achieve higher quality in the reconstructed frame when comparing with the BMA method. Moreover, subjective analysis reveal that this method recover the edges in the lost image. However, the method uses complex mathematical operations, which may not be feasible in real time applications.

3.2.2 Motion Vector Extrapolation

As mentioned before, the MC method is a simple way of concealment a missing frame using motion information from previously decoded frames. Although this method can provide an acceptable performance comparing with the FC, as shown in [76], more complex algorithms that use the assumption of motion continuity across consecutive frames are proposed to conceal, not only missing blocks, but also whole frames in compressed video.

To recover the entire missing frame a multi-frame motion vector averaging (MMA) algorithm was proposed in [85]. The method uses MVs of a few past frames (from $t-1$ to $t-L$), in order to estimate the forward MV for each pixel in the last received frame. Then, each pixel of the last frame is projected onto the missing one. More specifically, the method proposed in [85] recovers the missing frame at instant t using the following steps:

1. Generation of a MV history based on the previously decoded frames from $t-1$ to $t-L$, with $L \in [2,5]$, in order to understand which pixels were used to predict the ones of the frame at $t-1$.

2. The MV history is used to obtain a set of forward MV, from the frame $t-1$ to the missing one (at instant t). For each pixel, the forward MV is measured by averaging the MV history.

3. A spatial regularisation of the MV field is performed, by applying a median filter with a kernel size of 12, to approximate an edge-preserving regularisation [86].

4. The reconstruction of the missing frame is performed, at a half-pixel level, through the previously obtained forward MVs. The overlapped pixels are averaged, in order to obtain the final pixel value.

5. The unfilled regions are recovered, with the median value of a 7×7 window

6. Finally, the fully reconstructed frame at a half-pixel level is filtered using a mean filter with a 2×2 kernel, and down sampled.

This method usually produces relatively high quality in the reconstructed frame, but in low motion cases it behaves worse than the FC method under a few conditions, as indicated by the experimental results in [85]. This phenomenon happens because the oldest reference frames usually do not have MV similar to the missing one. Moreover, if L increases, the complexity of the proposed algorithm may not be useful to be used in error concealment, due to complexity and memory requirements to determine the MV history.

Motion vector extrapolation [87, 88] (MVE) methods are a simple but efficient way to reconstruct the MVs of a lost frame, and they are able to achieve good performance. In the method presented in [87], MVs are extrapolated from the last decoded frame (reference frame) onto the missing one. Firstly, each MB in the reference frame is projected to the missing one with its own direction given by the corresponding extrapolated motion vector (mv'). An example of motion vector extrapolation is shown in Figure 3.10, where the vectors of the frame $t-1$ are projected onto the missing frame at instant t. Note that, in the reference frame (f_{t-1}) the motion vectors may point to different frames

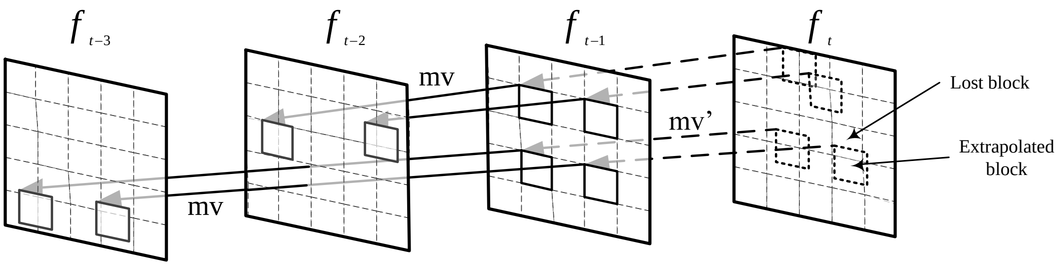


Figure 3.10: MVE extrapolation with the corresponding extrapolated macroblock, based on the method presented in [87].

(f_{t-2} and f_{t-3}), previously decoded, so MV extrapolation must take into account the difference between the original temporal distance and the projected one. Secondly, to improve the accuracy of MVE, the damaged MBs are divided into 4 blocks (8×8 pixels). Then, according to the matching area between the extrapolated MB and the damaged 8×8 block, the best MV is chosen. In order words, the MV for the each block is obtained from the extrapolated MB that achieves a higher number of overlapped pixels with the missing 8×8 block. Finally, the reconstructed MVs are used to fetch the missing pixels from the previously decoded frame (f_{t-1}). Although this method is able to overcome the disadvantage of incorrect MB displacement, the choice of the best MV is not very effective [89]. In [87] the MVE is used to extrapolate the MB of the reference frame. This results in overlapped regions, which are recovered using a weighted average of the multiple pixels candidates. The weight coefficients assume a maximum value in the center of the extrapolated MB, and a minimum value and the MB boundary.

Since the choice of the best candidate MV relies on the overlapped area of the extrapolated block with the lost region, the method may give erroneously estimated MV in large motion scenes. The example of Figure 3.11 shows a lost 8×8 block (in the centre of the figure) and four possible extrapolated MBs that can be used to estimate lost the motion information. As illustrated in the figure, although the largest overlapped area belongs to the MB_2, the overlapped pixels correspond to a small portion of the whole missing block. Moreover, the majority of the lost block's area was not covered by any extrapolated MB. Thus, motion recovery using only the information available in MB_2 will not properly characterise the motion information for all lost pixels.

To overcome the previous limitation and to achieve smoother and finer results, a pixel-based motion vector extrapolation (PMVE) algorithm was proposed in [89]. This motion vector extrapolation method is divided into two cases:

1. The pixels in the lost region that are covered by at least one extrapolated MB, such as circle points in Figure 3.11, are recovered using the average of the MVs of the overlapped MB.

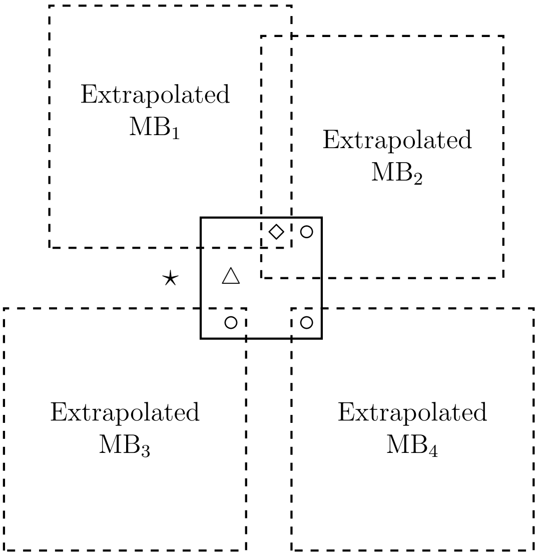


Figure 3.11: Pixel based motion vector extrapolation [89].

2. Those pixels which are not covered by any extrapolated MB, such as the triangle point in Figure 3.11, the MV of the co-located pixel in the previous frame is used.

Then, if the estimated MV is (MV_x, MV_y), then the estimation of the missing pixel, $p_m(x, y)$, is obtained by:

$$p_m(x, y) = p_r(x + MV_x, y + MV_y), \tag{3.6}$$

where p_r refers to the pixels in the previous frame. This pixel-based MVE provides similar results than the block-based MVE algorithm in low motion scenarios. Nevertheless, it is able to overcome the previous method in large motion scenes. An extension to the previous approach was considered in [89], by using the MVs of subsequent available frames. Thus, to estimate the lost motion information, forward and backward extrapolations are performed, in order to obtain two candidates to each lost pixel. Then, the lost pixel is estimated using the Equation 3.6, using the MV determined by the average of two MV candidates. The experimental results revealed that the bi-directional PMVE outperforms the isolated use of forward or backward PMVE. Moreover, the bi-directional PMVE achieves gains up to 1.85 dB and 1.32 dB comparing with the MMA [85] and MVE [87].

In [90] a hybrid motion vector extrapolation (HMVE) algorithm was proposed, to provide more accuracy to the estimated motion information. The method aims to improve the accuracy of the MVs of the PMVE method, by using the extrapolated vectors at the pixel and block level. Moreover, the proposed method discards the MVs that were wrongly extrapolated in order to obtain an accurate MV. In this concealment method,

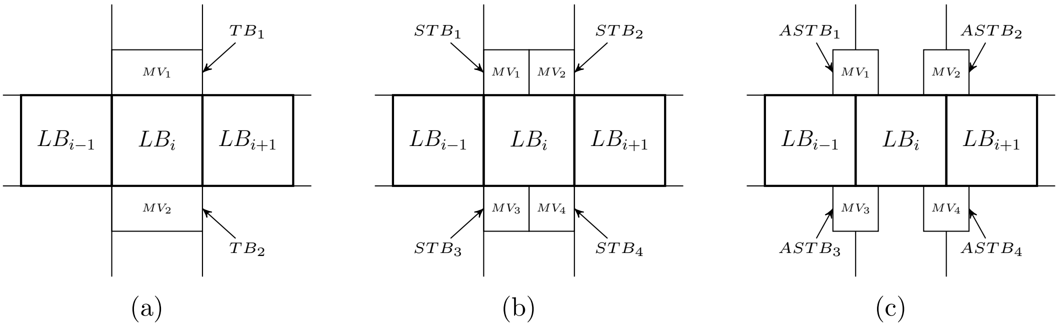


Figure 3.12: Average of the neighbouring MVs to conceal the MV of the centre 16 × 16 block: (a) using two vertical 16 × 8 neighbours blocks; (b) using four vertical 8 × 8 neighbour blocks; (c) using alternative four vertical 8 × 8 neighbour blocks as proposed in [91].

firstly a pixel based motion extrapolation is performed and the pixels is organised into three groups. Group A contains the pixels that are overlapped by at least one extrapolated block (both circle and diamond points in Figure 3.11). Group B includes the pixels that are not covered by any extrapolated block, but belong to a block that is covered, *e.g.*, in Figure 3.11 the triangle point was not covered by any extrapolated MB but it belongs to 8×8 block that is covered by 4 blocks. The remaining pixels are included in group C (star pixels in the figure). Then, through the block-based MVE method described in [87] two new MVs are determined: the first one corresponds to the MV of the most overlapped block with the current one, and second one is obtained by the weighted average value of the MVs corresponding to different overlapped blocks.

The HMVE method aims to find the more accurate MV from the MV set composed described above. To achieve this, the MV set is further refined using the distance between each pair of MVs, in order to discard the wrongly extrapolated MVs. Finally, the pixels of the lost regions (belonging to the groups A and B) are obtained through motion compensation, using the average of the remaining vectors in the MV set. The pixels of the group C, although the MV set is empty, the same process is applied using the MV of the co-located pixel in the previous frame. This algorithm is able to outperform the PMVE for different packet loss probabilities, as shown by the results in [90]. The method is used to conceal both I (INTRA) and P frames in video sequence compressed with the H.264/AVC, achieving more 1.09 dB (PSNR) in the erroneous frames.

Three techniques for error concealment are proposed for consecutive lost blocks, and their performances are investigated in [91]. The first technique uses the neighbouring MVs, to estimate the MV of the lost block (see LB_i in Figure 3.12). Firstly, an MV for the neighbouring blocks, referred as target blocks (TB), are estimated (if not available) using a block matching algorithm. Then, the MV for the lost block is obtained from the average value of the two vertical neighbour vectors, as shown in Figure 3.12(a). In order to increase the accuracy of the method, the target blocks are divide into two 8 × 8 small target blocks (STBs), as shown in Figure 3.12(b). However, this requires more computation, due to the motion estimation process. Thus, an alternative small target blocks (ASTB) are chosen, as shown in Figure 3.12(c). These can be located in the boundary of two adjacent lost blocks, which allows the estimated MVs to be used for recovery of two adjacent blocks. Therefore, the computational complexity remains the same as using the TBs as illustrated in Figure 3.12(a), but with higher accuracy.

The second technique presented in [91] also aims to reduce the computational time of boundary match algorithm, introducing the initial MV for the lost block, so the motion estimation using a boundary match algorithm can use a reduced search range. The third technique aims to recover those areas where no neighbour blocks are available to perform the motion estimation. Therefore, the optical flow of the neighbouring blocks is calculated using the Horn-Schunch algorithm [92], and the optical flow information is used in the concealment. In comparison, the boundary matching was the most successful one, except when the loss of multiple lines occurs, in which optical flow algorithm achieves higher quality.

3.3 Concealment approaches for 3D video

In the previous sections different concealment algorithms to cope with missing information in 2D video were described and compared. Although these methods can be used in 3D video, their performance can be further increased using additional information, such as the relation between the 3D viewpoints [93]. In this section different algorithms are described for concealment of missing information in stereoscopic video.

A full frame loss algorithm for stereoscopic video was proposed in [94]. In this study the full frame loss concealment algorithm proposed in [87] is extended to stereoscopic video. Two methods to conceal the missing frames are proposed. These methods used the relation between the number of motion compensated block and disparity compensated ones in the temporally closest frame. One the one hand, if in a given sequence more motion compensated blocks are used, the right view frames of stereoscopic video are concealed, as in the monoscopic algorithm [87]. On the other hand, if more disparity compensated block are used, the frames are concealed using the disparity vector of the reference right view frame. The results show that this method is able to adapt to a given stereoscopic sequence. It chooses the best method, achieving an acceptable subjective video quality.

As shown before, the concealment methods for 3D video may use the techniques proposed for 2D video, improving the concealment performance by introducing new features. Two concealment methods for 3D video was proposed in [95], to cope with losses at the

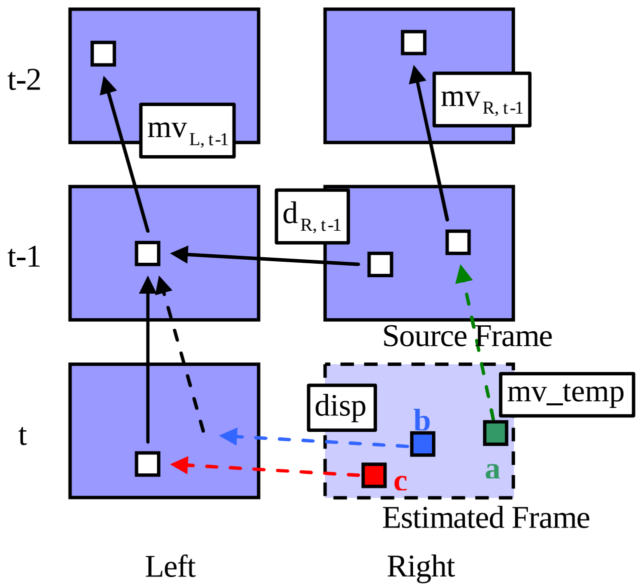


Figure 3.13: The projection methods used in [95] for the estimation of the right frame at time t.

block and pixel level using a combination of the motion and disparity vectors from the neighbouring frames. The methods aim to recover the missing frames in the dependent view using both the temporal extrapolated MVs and the disparity vectors (DVs). In Figure 3.13 the left and the right frames are represented, and the frame at instant t is missing. The method recovers the MVs for the missing frame using one of the following techniques, represented in Figure 3.13 by the block 'a', 'b' and 'c' respectively:

1. For the block 'a' the MV $(mv_{R,t-1})$ of the source frame $(f_{R,t-1})$ are temporally extrapolated as in [87].

2. For the block 'b' the DV $(dv_{R,t-1})$ are projected into the missing frame using the extrapolated version of the MV $(mv_{L,t-1})$ corresponding to the block, in the left frame at the instant $t-1$, referenced by the $dv_{R,t-1}$.

3. The block 'c' illustrates a particular case in which the block referenced by the $dv_{R,t-1}$ corresponds to an INTRA macroblock. Therefore, a zero MV is considered and the method described in 2 is used.

Then, the motion and disparity field are filtered and combined to generate a unique vector field. Then, the missing frame is reconstructed through motion/disparity compensation. Finally, some of the unassigned pixels may be filled by a post 9×9 median filter or, when this does not work, using a zero motion vector. The concealment algorithm at the pixel level achieves higher performance than the same algorithm at the block level. However, it takes approximately ten times longer to conceal the missing frames [95].

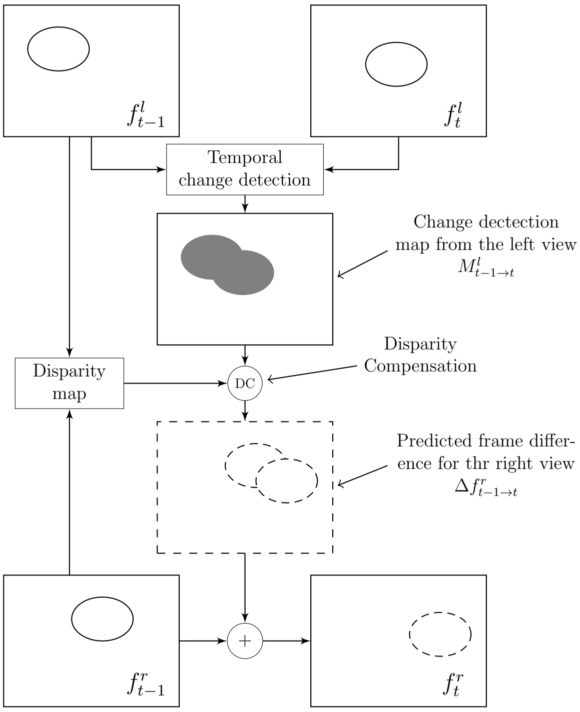

f^l_{t-1}
f^l_t
Temporal change detection
Change dectection map from the left view $M^l_{t-1\to t}$
Disparity map
DC
Disparity Compensation
Predicted frame difference for thr right view $\Delta f^r_{t-1\to t}$
+
f^r_{t-1}
f^r_t

Figure 3.14: Illustration of the disparity-based frame loos concealment method proposed in [96]

The above mentioned concealment method relies on the assumption of a constant translational motion in the right view frames, which may fail and degrade the subjective and objective quality of the reconstructed frame. The work presented in [96] aimed to propose a frame loss concealment to recover the missing frame based on the disparity and the difference of the temporally adjacent frames in the error-free view. The main idea of the proposed method is illustrated in Figure 3.14, where a missing frame at instant t in the right view, f^r_t. The underlying principle is to project the temporal displacement data (*i.e.*, frame difference) from the error-free view, $\Delta f^r_{t-1\to t}(x,y)$, into the other view to reconstruct the missing frame. As shown in the figure, the temporal change detection is measured from adjacent frames in the left view, f^l_{t-1} and f^l_t, and marked in the change detection map, $M^l_{t-1\to t}$. The disparity is estimated using a global search followed by refinement at the instant $t-1$. Then, the relevant temporal changes are disparity compensated, in order to obtain the frame difference for the right frame, $\Delta f^r_{t-1\to t}(x,y)$. Finally, the reconstruction of the missing frame is performed as follows:

$$f^r_t(x,y) = f^r_{t-1}(x,y) + \Delta f^r_{t-1\to t}(x,y). \tag{3.7}$$

The proposed method is compared with the methods [94] and [95], achieving consistently higher objective and subjective quality. Although this method may be applied to a burst of lost frames, it was not taken into account in the experimental setup. Moreover, it does

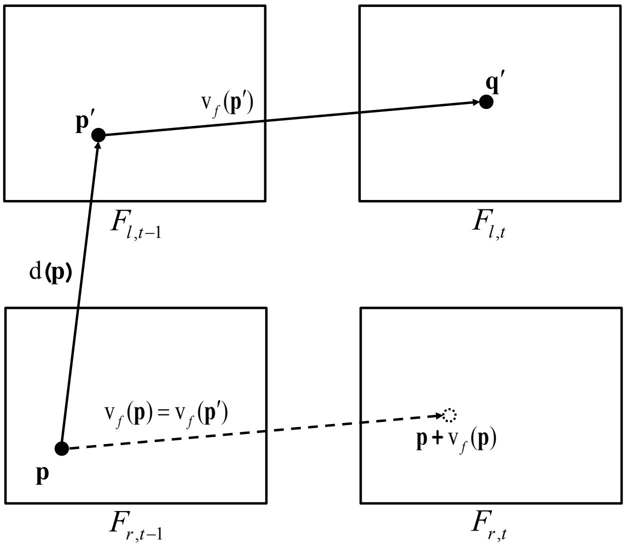


Figure 3.15: Motion projection from the left to the right view to conceal the missing right frame at the instant t ($F_{r,t}$) [97].

not deal with illumination changes between views, which may happen in stereoscopic video.

Another method based on the motion similarities between the left and right view of the stereoscopic video is evaluated in [97]. Figure 3.15 represents the iterations to recover the missing frame of the right view ($F_{r,t}$) at the instant t. As shown in the figure, based on the motion information of the error-free frame $F_{l,t}$, a set of forward MVs are generated for the pixels within the frame $F_{l,t-1}$, *e.g.*, $v_f(p')$. For each INTRA block in $F_{l,t}$, the forward motion vector is set to a zero. Then, the intensity difference of the pixel p' is defined as:

$$\epsilon(p') = F_{l,t}(q') - F_{l,t-1}(p'). \tag{3.8}$$

The disparity map (referred to as d) is estimated through a block matching algorithm using the stereo pair at instant $t-1$. Then, the missing pixels are reconstructed by,

$$F_{r,t}(p + v_f(p)) = F_{r,t-1}(p) + \epsilon(p), \tag{3.9}$$

assuming,

$$\begin{aligned} p' &= p + d(p), && (3.10)\\ v_f(p) &= v_f(p'), && (3.11)\\ \epsilon(p) &= \epsilon(p'). && (3.12) \end{aligned}$$

Finally, the empty regions corresponding to occlusions or unfilled regions are concealed

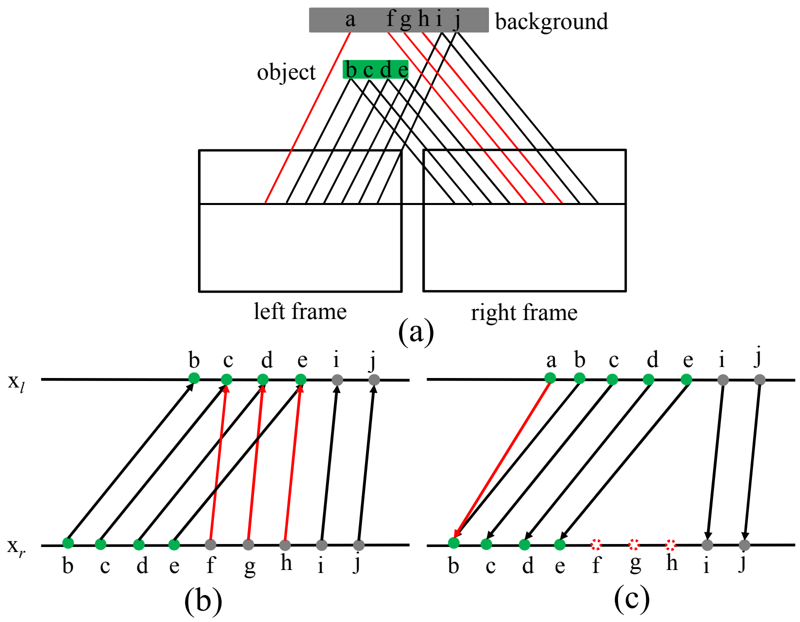


Figure 3.16: (a) Projection of object and background points onto left and the right frame; (b) Matching from the right frame to the left frame, 'f', 'g', and 'h' are incorrectly mapped to 'c', 'd', and 'e'; (c) Matching from the left frame to the right frame, 'a' is incorrectly mapped to 'b', and no pixel is mapped to 'f', 'g', or 'h' [98].

using a spatial or temporal replacement. The value of temporal intensity difference (*e.g.* $\epsilon(p)$ in Figure 3.15), of the neighbour pixels is measured, to decide which replacement should be applied. This method is compared against the pixel-based approach presented in [95] and the [96]. The results shows the advantage of the proposed method, not only in the missing frames, but also in the subsequently affected ones. However, the method performs a rough estimation of the occluded pixels, therefore, the motion information is not shared between views for all non-occluded pixels. Since a spatial replacement or a zero MV is used in the occluded pixels, the performance of the method decreases.

The same approach is also used to conceal the missing right colour frames of the stereoscopic plus depth video [98], however, the disparity map is obtained using the available depth maps and the scene acquisition information. Moreover, it was proposed an occlusion detection algorithm. Figure 3.16 represents a matching example, with some occluded pixels. In the figure x_l and x_r corresponds to left and right frames respectively. Figure 3.16(a) shows that object points 'b', 'c', 'd', and 'e' and background points 'i' and 'j' are visible in both left and right frames. However, background points 'a', 'f', 'g', and 'h' are occluded. Figure 3.16(b) illustrates the matching of pixels from the right frame to the left one. Although pixels 'b', 'c', 'd', and 'e' are correctly matched, the occluded pixels 'f', 'g', and 'h' are incorrectly matched. Figure 3.16(c) represents the inverse matching. Similarly, pixels 'a' is incorrectly matched to 'b' and no pixel in x_l is matched to 'f', 'g', or 'h' in x_r. Therefore, a pixel is considered as occluded if one of the following conditions

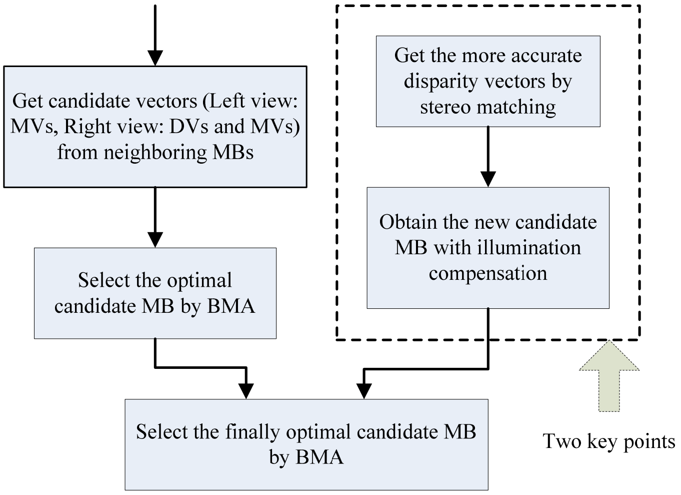

Get candidate vectors (Left view: MVs, Right view: DVs and MVs) from neighboring MBs
Select the optimal candidate MB by BMA
Get the more accurate disparity vectors by stereo matching
Obtain the new candidate MB with illumination compensation
Select the finally optimal candidate MB by BMA
Two key points

Figure 3.17: Flowchart of the error concealment algorithm for stereo video based on the illumination changes between views proposed in [99].

is true: it shares the matching pixel with other pixels; it is matched by multiple pixels; it is not matched by any pixel. Using this algorithm, instead of conceal the occluded pixels with the method presented in [97], the method proposed in [98] uses the temporally extrapolated vectors within the affected view (right view).

In 3D video, illumination changes may occur due to imperfect calibration and variations of the camera position and direction. The illumination changes between different views of the 3D video can cause quality degradation in the concealed frame, when inter-view dependencies are used to obtain the reconstructed frame from the neighbouring view. The method proposed in [99] mainly deals with the problem caused by illumination changes between two views during the inter-view error concealment. The method aims to recover the missing MB using the available information in the neighbour MBs. Figure 3.17 shows the flowchart of the method, and as shown it is divided in two parts. Firstly the traditional method based on BMA is performed to get an optimal candidate MB. Secondly the accurate disparity vector of the lost MB is measured by stereo matching, and the illumination compensation on the MB which should be copied from the left view is performed.

The disparity map measurement uses the normalised cross correlation (NCC) as the similarity measure, because the NCC is insensitive to the illumination changes [100]. The NCC is evaluated in a small window of pixels, which are available in the neighbourhood of the lost MB. Then, a simple illumination compensation value (ICV) is used to measure the illumination change between two views. The ICV value is defined as the difference between the average illumination values of the corresponding pixels of two views [47], and is measured as follows,

$$ICV = \left(\sum f_R(x,y) - \sum f_L(x + D(x,y), y)\right) \Big/ N \, , \tag{3.13}$$

where f_R, f_L correspond to the right and left frames respectively, N is the number of pixels in the matching window and D is the disparity value. The ICV value is used to compensate the pixels values of the left view MB used as reference. Finally, as shown in Figure 3.17 the best temporal/spatial candidate is selected using a BMA algorithm, *i.e.*, using a distortion based algorithm. This method deals with illumination problems, providing relevant knowledge to other concealment methods based on inter-view similarities. However, if no neighbour blocks are presented, there is no information to determine the ICV value, so this method fails.

3.4 Summary

This chapter presented an overview of the different error concealment methods, proposed over the last years. Error concealment methods for 2D video provide relevant insights to develop novel techniques to recovery losses in 3D video. The contributions of this research work are focused in error concealment techniques for 3D video (stereoscopic video), therefore, the error concealment methods for depth and multi-view were not described.

Different error concealment methods for 3D video described in Section 3.3 are based on traditional techniques applied to 2D video. Based on this, firstly the subjective impact of three frame loss concealment methods at frame lever are discussed in the Chapter 4. Subsequently, in the Chapter 6 a novel error concealment technique is proposed using a combination of a temporal 2D video concealment technique (*i.e.*, MVE), enhanced with inter-view dependencies, so an accurate motion/disparity field for the missing frame can be achieved.

Chapter 4

Subjective evaluation of temporal distortion in 3D video

Despite the different characteristics of the existing formats and their associated compression techniques [11], the quality of the user experience (QoE) provided by 3D video delivery services consistently depends on the need for adaptation of compressed streams to the diverse networking technologies, transmission errors and packet losses. For instance, rate reduction of compressed streams based on frame skipping has been used in the past as a network adaptation approach to match stringent bandwidth constraints in mobile communications [101]. Moreover, in video transmission over error prone channels, packet loss may occur, causing video quality degradation, mainly due to frame loss. A detailed discussion about 3D video transport over different communications networks and associated problems can be found in [7].

In this chapter, a subjective evaluation study is presented, in order to find the objective factors that have the highest influence in the perceived quality of 3D video, affected by temporal distortions caused by missing frames at the receiver. Subjective evaluation is a relevant indicator for 3D video quality, since it depends on several perceptual attributes such as image quality, perceived depth, naturalness and eye strain, among others [102]. Two different cases of practical interest are investigated: the case of regular frame skipping, due to constrained encoding and/or transcoding, and the case where random losses occur during transmission in the network. In this context three frame loss concealment methods at the frame level are proposed and assessed through subjective testing, in order to evaluate the perceptual impact of error concealment in the temporal domain. The rest of this chapter is structured as follows: Section 4.1 presents an overview of the current state-of-the-art in perceptual evaluation of 3D video sequences. Section 4.2 describes the experimental setup, regarding regular and random error pattern generation and the investigated frame loss concealment methods. Section 4.3 discusses the subjective quality assessment procedure that was used is this work. Section 4.4 presents the experimental results of this study, as well as a discussion on the most important issues regarding viewers' evaluation of the tested impairments and the used concealment conditions.

4.1 Related work

In the process of depth perception, the human visual system (HVS) is responsible for acquiring two slightly different views of the same scene by using two eyes located horizontally apart. Additionally, several other monocular clues are also relevant to perceive depth information in stereo video, such as occlusion, relative object size, motion parallax, lighting and shading. In general, viewers' opinions are favourable to 3D video, but the stereoscopic system must provide a comfortable immersive experience, free of artifacts that may be introduced by any component of the 3D communication system. For this reason, in the last few years, several research works have addressed 3D video quality issues. For instance, error concealment techniques, objective metrics for measuring subjective quality and numerical relationships between 3D video quality and 2D content features are among those aspects of highest importance in current 3D video communications [103–105]. However, many of the 3D quality related aspects are still quite open for research, because most of the 3D perceptual experience is brain-driven, which makes it difficult to characterise and to find appropriate mathematical models. This is the case of the impact of temporal artifacts on the perceptual quality of 3D video addressed in this chapter.

The effects of temporal artifacts in the perceived quality have been investigated in the past for 2D video signals [106, 107], where the viewers' opinion was related to the motion of the scene, characterised by different descriptors. Experimental results showed that the perceived quality depends on the levels of strength, duration and distribution of the temporal impairments. For instance, the impact of frame decimation depends on the original frame rate of the sequence. For a given frame rate, the perceived quality does not always decrease for higher motion magnitudes, suggesting that other factors may affect perception of temporal artifacts.

In the literature, there are several studies about perceptual quality assessment for stereo pairs, proposing objective quality metrics for 3D images. As for stereoscopic video the perception evaluation for 3D images must take into account several new elements, when compared with the case of 2D content. Some previous studies have evaluated the impact of reducing the spatial resolution for one view of the stereo pair [108–110]. It has been demonstrated that mixed-resolution stereo image sequences can still provide acceptable image quality. In [111] and [112], the authors combine monoscopic image quality metrics and depth information, in order to propose a new method to evaluate the quality of stereo pairs. For instance, the method proposed in [113] takes into consideration camera information and image resolution, whilst in [114], the authors use computational models of the HVS to develop a new stereoscopic image perceptual quality metric.

The response of the HVS to mixed-resolution stereo video-sequences, in which one view was spatially or temporally low-pass filtered, was recently investigated in [115]. The effects on the perception of the overall quality, sharpness and depth sensation were evaluated. Experimental results showed that the overall sensation of depth was unaffected by the reduction of spatial resolution in one of the views. The temporal artifacts were only investigated for a regular frame drop using 1/2 and 1/4 temporal resolution, field averaging and frame drop-and-repeat as concealment methods. These temporal artifacts were rated as unacceptable, regarding the overall quality and image sharpness, but showed a small impact on depth perception.

Recent works have used auto-stereoscopic displays for subjective evaluation experiments [102, 116, 117]. However, it is a well know fact that this display technology is not yet fully mature and that 3D perception in such type of displays is highly dependent on the viewer relative position [20]. More mature technology such as 120Hz displays with synchronised shutter glasses is already in the consumer market, which makes it more relevant for subjective quality assessment.

4.2 Frame loss concealment

A subjective evaluation framework has been defined to assess the effects of temporal distortion caused by frame loss on the perceived 3D video quality. It comprises two models that simulate frame loss and three concealment methods at the decoder. Frame loss is assumed to have two sources: regular frame dropping, as result of frame skipping encoding/transcoding schemes, and random frame loss, resulting from transmission errors and/or congested packet networks. In the latter case, each single packet is capable of carrying either complete frames or most of their coded data. Thus, each lost packet results in a single frame loss. Both regular frame skipping and random frame loss lead to missing frames at the receiver, which are recovered by implementing a loss concealment strategy for the whole frame.

The frame loss concealment methods used in this chapter are representative of the main classes of concealment methods to recover from frames loss. Many variants can be derived from the three methods (Frame-Freeze, Double-Freeze and Base-Copy) investigated in this study. As an example, Frame-Freeze can be done with or without motion compensation, but these are two variants of the same method. More details about each method are given in the next sub-sections.

4.2.1 Frame loss models

Regular loss

In the frame skipping model, three patterns, with two variants, were used for quality evaluation. The skipping patterns are defined as 1×2, 2×3 and 3×4, which corresponds to drop every other frame, 2 out of 3 and 3 frames out of 4, respectively. For each of these patterns, two different scenarios were implemented: in the 1×100 variant, frame skipping occurs during the entire sequence (100%), while in 1×50, only the last half of the sequence is affected by frame skipping (50%). These two scenarios were used to evaluate possible perceptual differences between constant frame skipping and the transition from high to low temporal rates in stereoscopic view.

Random loss

The impact of random occurrence of frame losses was investigated using a frame loss model based on a two-state Markov chain. A finite state Markov chain (Figure 4.1), characterised by the good (G) and bad (B) states and transition probabilities p and r, was implemented to generate the loss patterns used to induce temporal distortion by dropping the corresponding frames. Such type of loss models are commonly used to simulate data loss errors in different scenario applications [118].

The steady state probabilities, π_G and π_B, are defined as the probabilities of the chain to be in either the G or B state, respectively, after an infinite number of steps. These probabilities are independent of the initial state. Transitions between states are defined as conditional probabilities $p = P(B/G)$ and $r = P(G/B)$, that depend on the previous state. Each state of the Markov chain is associated with an error generating process, where the error rate associated with the G state is defined as $1 - k$ and that of B state is $1 - h$. In the simplified Gilbert model, commonly used in error/loss simulation, the G state is error-free (*i.e.*, $k = 1$), then the average error rate is given by $p_E = (1 - h)\pi_B = (1 - h)p/(p + r)$ [119].

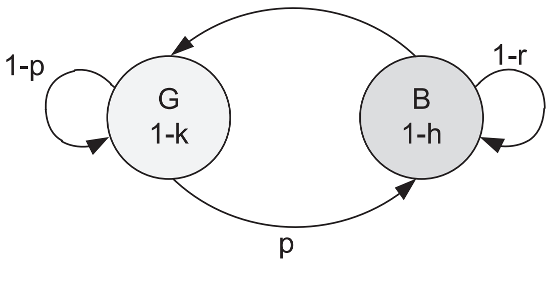


Figure 4.1: The two-state Markov chain used to simulate random loss.

Therefore, in the Gilbert model used in this subjective evaluation, the good state G corresponds to a successful reception of a frame, whilst the bad state B represents a failure in the reception of a video frame. The output of this model is a frame error trace file, where the number 1 means a lost frame and 0 represents a successfully received frame. The model parameters used to achieve the desired overall error probabilities in trace files were set in the experimental process.

4.2.2 Concealment methods

Three frame loss concealment methods were used to assess the subjective impact of temporary loss of depth information. As pointed out before, the objective of this study was to evaluate the impact of missing frames on the perceived depth. In order to isolate the effect of temporal 3D distortion from classic problems of 2D video transmission, the base view is assumed to be transmitted over a high priority channel, that guarantees an error free transmission of this 2D video sequence. Since the auxiliary view is transmitted over an error-prone channel, only depth information is affected when frames in the auxiliary view are discarded. Three different frame loss concealment methods were used to evaluate the quality degradation experienced by users, in the presence of temporary loss of the auxiliary view:

- The first concealment method is named Frame-Freeze (FF) and consists in copying the last received frame of the auxiliary view into the temporal instant where the lost frame occurred in the same view (Figure 4.2). FF may introduce some interview (disparity) distortion, since the reconstructed sequence is composed by base and auxiliary frames that are unpaired in the original sequence.

- The second method, named Double-Freeze (DF), both the base and auxiliary views' frames are copied into the temporal instants where frame loss occurred. This concealment strategy corresponds to freezing the last 3D image received without errors (Figure 4.3). This means that DF freezes two stereo views that were paired in the initial sequence.

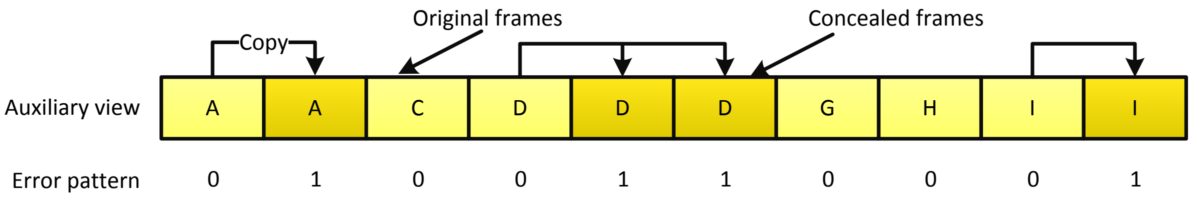


Figure 4.2: Frame-Freeze concealment method (FF).

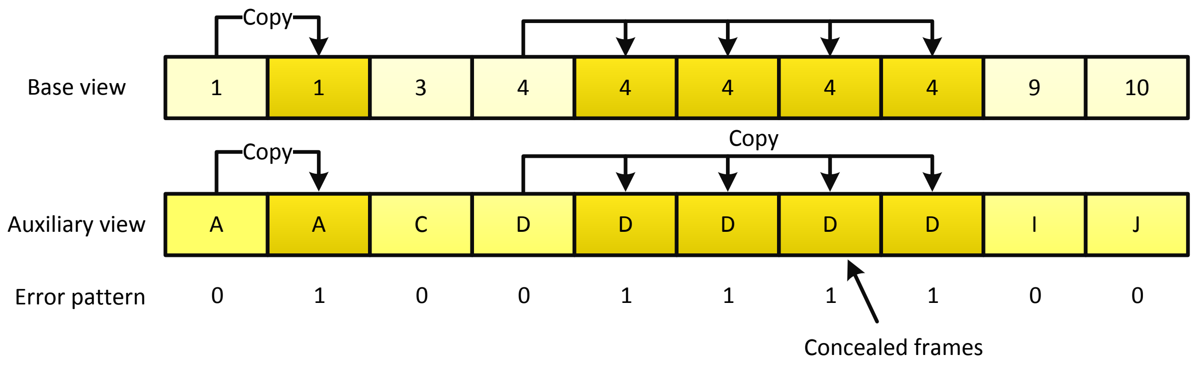


Figure 4.3: Double-Freeze concealment method (DF).

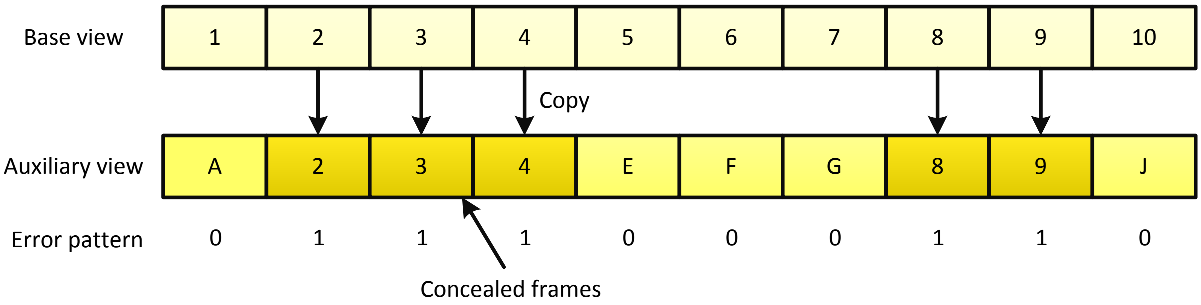


Figure 4.4: Base-Copy concealment method (BC).

- The third method, named Base-Copy (BC), copies the temporally co-located frame from the base view to the auxiliary view, in order to fill the corresponding missing frame at the relevant time instants (Figure 4.4). In practice, the BC method switches the sequence from 3D to 2D during the error period, by presenting the same frame to both eyes of the viewer.

4.3 Evaluation of perceived 3D video quality

The perceived 3D Video quality was evaluated through subjective testing. The aim is to relate user perception with temporal distortion, content's depth characteristics, frame loss statistics and the used concealment method. These relations should reveal subjective factors that may be taken into account in future practical implementations. Seven stereoscopic video sequences, with different characteristics, were subjected to frame losses in the auxiliary view, according to the previously described regular and random loss schemes. The test environment and the assessment methodology used in the subjective evaluation process were carefully defined in order to meet the standard requirements [120–122]. The three frame concealment methods were tested for different error probabilities and average loss error burst.

Table 4.1: 3D video sequences used for subjective evaluation.

Name	Content description	Disp.
Champagne Tower	A woman stands next to a cup pyramid and handles one of the cups; moderate motion and high depth.	21.5
Pantonime	Two clowns moving around near a box; moderate motion and moderate/high depth.	17.7
Kendo	Two men practice kendo with spectators on the back; moving camera; high motion and moderate/high depth.	16.7
Balloons	A man entering on a big balloon while moving; moving camera; intense/complex motion and moderate/high depth.	16.0
Jungle	A toy plane shaking while another toy moves under it; moderate motion and moderate depth.	13.7
Uli	Two men talking; low motion and low depth.	10.4
Dog	A woman playing with a dog in front of a curtain; moderate motion and low depth.	8.8

4.3.1 3D video sequences

Seven different 3D video sequences with duration of 12 seconds and resolution of 1024 × 768 @ 25Hz were used in the whole evaluation process. No audio track was used in these subjective tests. All sequences exhibit different features, regarding scene content, motion complexity and disparity intensity. The main characteristics of the test sequences are presented in Table 4.1 and the first frame of each one is shown in the Appendix A. The disparity (Disp) was obtained as a normalised sum of the absolute pixels displacement between both views of the stereo sequence. This is a rough measure of the sequence disparity, but the subjective tests show that there exist a correlation between Disp and the 3D perceived quality (Section 4.4). Two of the test sequences (Uli and Jungle) were only used during the demonstration period, described in Section 4.3.3, so the subjective evaluation is not affected by this period. Thus, only the remaining five sequences were used for the subjective evaluation of the 3D temporal distortion.

4.3.2 Test environment

The subjective quality assessment experiments were conducted in a test room conforming to the viewing conditions defined in BT.500 [120]. The background lighting conditions used daylight colour temperature with an intensity of 200 Lux. The room is also equipped with white walls and an adequate sound insulation system. A workstation was used to

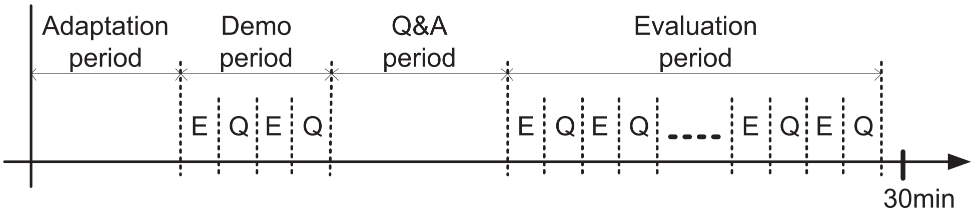


Figure 4.5: Time pattern of a subjective evaluation session.

reproduce all the test sequences, as well as to register the subjective test results. A 22-inch computer LCD monitor with a frame rate of 120Hz, a native resolution of 1680 × 1050 pixels and a dot pitch of 0.282 mm was used to display the test sequences. The viewers used liquid crystal shutter glasses, while seating in an arm tray chair. The viewers chair was placed at a distance of eight times the physical height of the picture from the screen, according to the preferred viewing distance (PVD) table in [120], but the participants were allowed to adjust their position to the most comfortable viewing distance. Each sequence was evaluated immediately after the observation, saving the result on a local database through a proper interface.

4.3.3 Subjective assessment methodology

The Single-Stimulus Absolute Category Rating (ACR) method was used in these experiments [121]. The processed video sequences (*i.e.*, with concealed frames) were divided into three groups, to avoid exposing the viewers to a session that would last more than half an hour [120]. Each viewer evaluated only the sequences of one group, presented in random order, but each sequence was evaluated by the same number of viewers. Each test sequence was played for 12 seconds and afterwards the viewer was asked to judge the 3D perception using a five level scale: 1 (Bad) to 5 (Excellent). The voting period for each sequence was not time-limited. After each evaluation, the observer was asked to click a button to proceed to the next sequence.

Each evaluation session was comprised of one continuous test session, divided into four parts, as shown in Figure 4.5. During the adaptation period, the observers stereopsis was first assessed using a random dot stereogram test, as suggested by ITU recommendation for stereoscopic subjective assessments [122]. All participants reported to have a normal acuity and all passed the random dot stereogram test. Furthermore, some synthetic and natural 3D sequences were presented to all observers, prior to the testing period, in order to allow them to adapt to stereoscopic viewing.

In the demonstration period (Figure 4.5), two sequences with the temporal artifacts used in our study were presented to each viewer, in order to avoid the influence of any surprising effect in the perceptual evaluation and also to allow an adaptation to the assessment method. During this period, the viewer simulates the evaluation process by experiencing the 3D viewing (E) and assessing the quality (Q). A period for questions and answers (Q&A) was then allowed, during which the viewer could ask some questions regarding the test. Finally, in the assessment period the participants observed each sequence (*i.e.*, E periods) and rated the corresponding quality (*i.e.*, Q periods), according to the previously described methodology.

A total of 59 male and 27 female viewers participated in the subjective evaluation, aged from 15 to 55 with an average age of 28.4 years old. Each 3D video sequence was evaluated by a random group of 16 to 19 observers. Before each test, a set of instructions were provided to users in a written form, and then they were given the opportunity to ask any questions regarding the evaluation process. The total number of viewers per sequence conforms with the minimum value of 15, specified in the international recommendations [120]. The participants were randomly recruited among general public, mainly composed by students and faculty staff. None of the participants had previous experience in a subjective video quality assessment procedure.

4.4 Experimental results

The experimental procedure was divided into two sets of subjective 3D quality evaluation tests. The first set of experiments was conducted to assess how the user 3D experience is affected by regular frame skipping, using different frame loss concealment methods, whilst the second set of experiments was carried out with random frame loss using the same concealment methods. Since these sequences present different levels of motion and disparity, the perceived quality was evaluated with the mean opinion score (MOS) of the original sequences as a function of the corresponding disparity. The plot in Figure 4.6 shows that viewers tend to give higher scores to sequences with higher disparity. This means that viewers consider depth (*i.e.*, disparity) a relevant factor in 3D video quality. The higher the disparity, the better is the viewing experience of the 3D content, in the absence of any temporal artifacts.

4.4.1 Regular frame skipping

The first experiments use image sequences with a certain percentage of regular frame loss in the auxiliary view, namely 1×2, 2×3 and 3×4. Figures 4.7 and 4.8 show the resulting MOS for the frame loss concealment methods previously described, namely FF, DF and BC. In both figures, the MOS value was obtained for different frame loss patterns (from 1×2 to 3×4). Figure 4.7 shows the results for temporal impairments during the second half of the sequence (1×50), while the results shown in Figure 4.8 were obtained for a

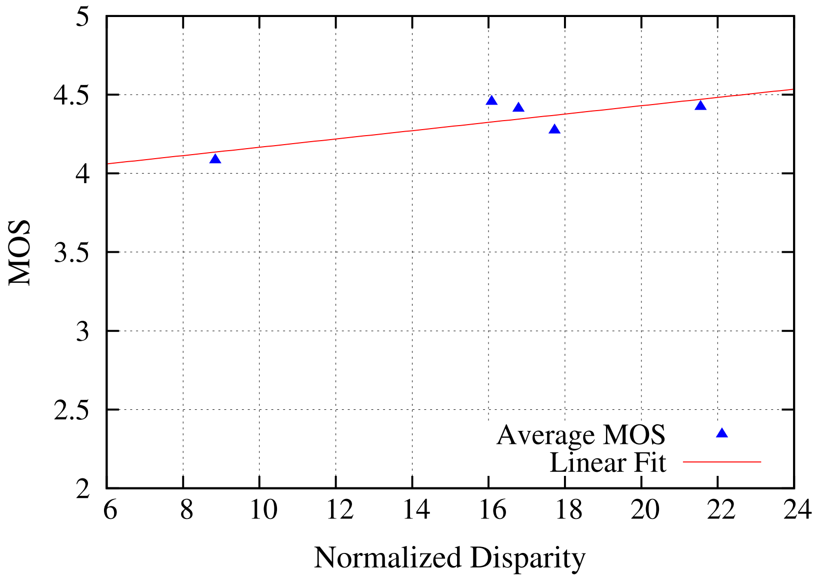


Figure 4.6: MOS versus disparity for error-free sequences.

temporal error that spanned the whole duration of the 3D viewing experience (1 × 100).

A common result in both Figures 4.7 and 4.8, is observed when the frame loss rate increases from 1 × 2 to 3 × 4. In this case, a MOS reduction is obtained for both FF and DF methods, while for the BC method the MOS increases. This effect can be explained due to the nature of the concealment methods. While both FF and DF replicate the previous frame in the auxiliary view (frame freeze), the BC method temporary converts the video sequence from 3D into 2D, using the base frame in both views. The increasing MOS, obtained with the BC method, reveals a user trend to find temporary switch from 3D to 2D video perceptually more pleasant than using other 3D concealment methods, for longer periods of missing frames.

Comparing the results of Figure 4.7 and Figure 4.8, one can observe that for all concealment methods and frame loss rates, when the frame loss occurs only in the last half (1 × 50) of the sequence (Figure 4.7), the MOS is on average about 10% higher than in the case of 1 × 100, when the loss occurs for the entire sequence (Figure 4.8). This seems to indicate some memory effect, since users give higher scores to sequences when errors occur after a period without any artifacts. In both cases, the FF concealment method results in higher average scores than the others, but the MOS value decreases as the percentage of loss increases.

The previous observation about switching from 3D to 2D is confirmed by the results of Figure 4.9, where it is can observe that for all sequences, when the frame loss rate increases and the video sequence is shown in 2D for a longer period (*e.g.*, 3 × 4), the MOS also increases. It is worthwhile to note that MOS is lower for sequences with higher disparity (see Table 4.1). This means that users perceive higher 3D quality after frame

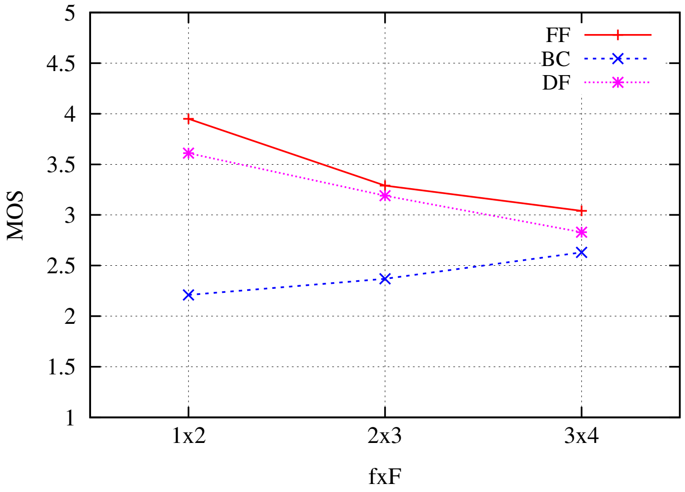


Figure 4.7: MOS versus percentage of regular frame loss (fxF) for each concealment method (1 × 50).

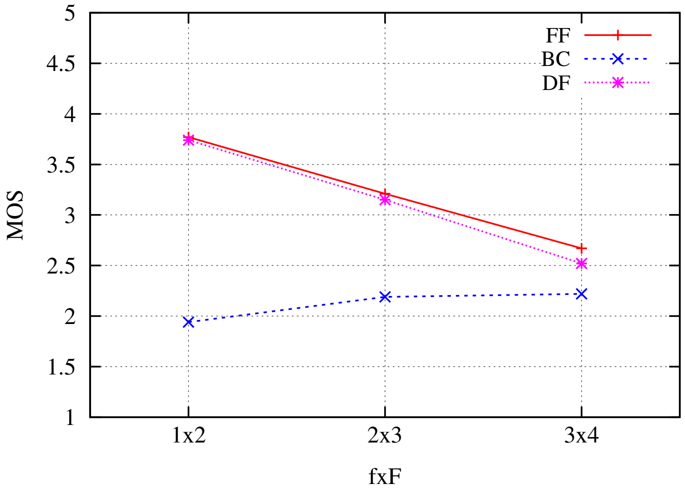


Figure 4.8: MOS versus percentage of regular frame loss (fxF) for each concealment method (1 × 100).

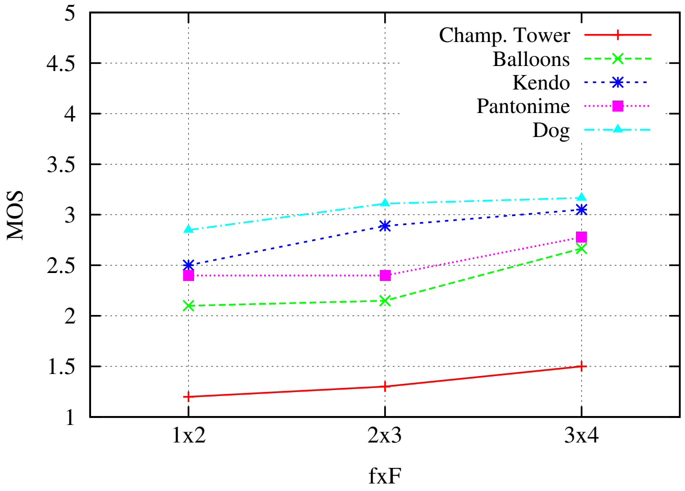


Figure 4.9: MOS versus percentage of regular frame loss (fxF) in 50% of the sequence (BC method).

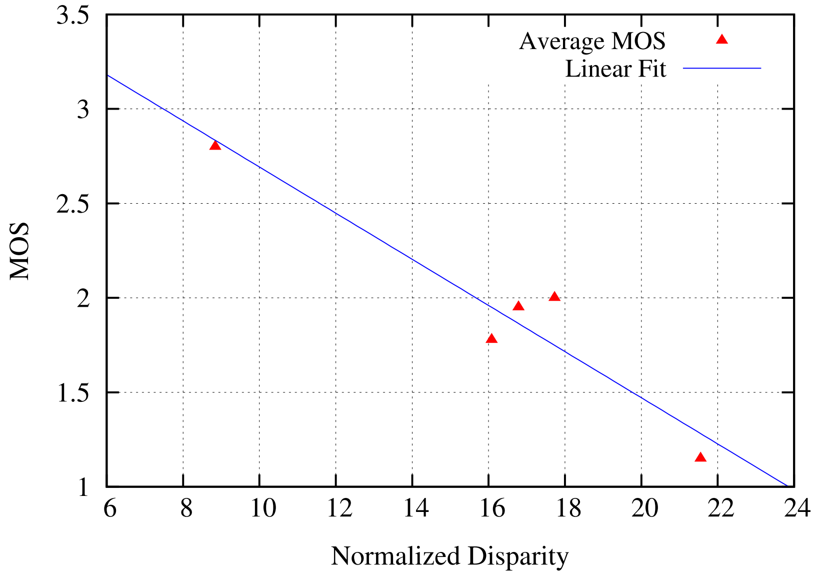


Figure 4.10: Disparity versus MOS for BC method in 50% of the sequence and 1 × 2.

loss concealment, when the original video sequences have lower depth (*i.e.*, disparity) information, *i.e.*, sequences perceptually closer to 2D than 3D. The results of Figure 4.9 also show that observers prefer to keep the same format, either 2D or 3D, rather than continuously toggle between them, as it happens in the case 1 × 2.

Figure 4.10 shows the average MOS obtained for each sequence, for the three error concealment methods, versus its corresponding disparity, for a rate loss of 1 × 2 in half of the sequence. This Figure clearly confirms that MOS decreases as the disparity increases. As pointed out before, the MOS for sequences that have higher disparity is more penalised when they are affected by errors. This shows that losing 3D perception has a stronger subjective impact when the 3D perception of the original sequence is higher.

4.4.2 Random frame loss

In the second set of experiments, the impact of random frame loss in the perceived quality of 3D video was investigated, for the same five stereoscopic video sequences, at different frame loss ratios (FLR). The missing frames were also reconstructed using the concealment methods described in Sub-section 4.2.2. The statistics of the loss patterns generated by the Markov model presented in Sub-section 4.2.1 are shown in Table 4.2.

Table 4.2 shows the average percentage of frame loss and error burst used to simulate the frame loss process. This process is characterised by the number of bursts (N) occurring in the sequence and their minimum (Min), maximum (Max) and average (Avg) length. While the average burst length is kept roughly constant for all FLRs, the number of error bursts and their size variability is quite different. The longer burst (size=33) was

Table 4.2: Statistics of loss frame patterns used in the random frame loss subjective tests.

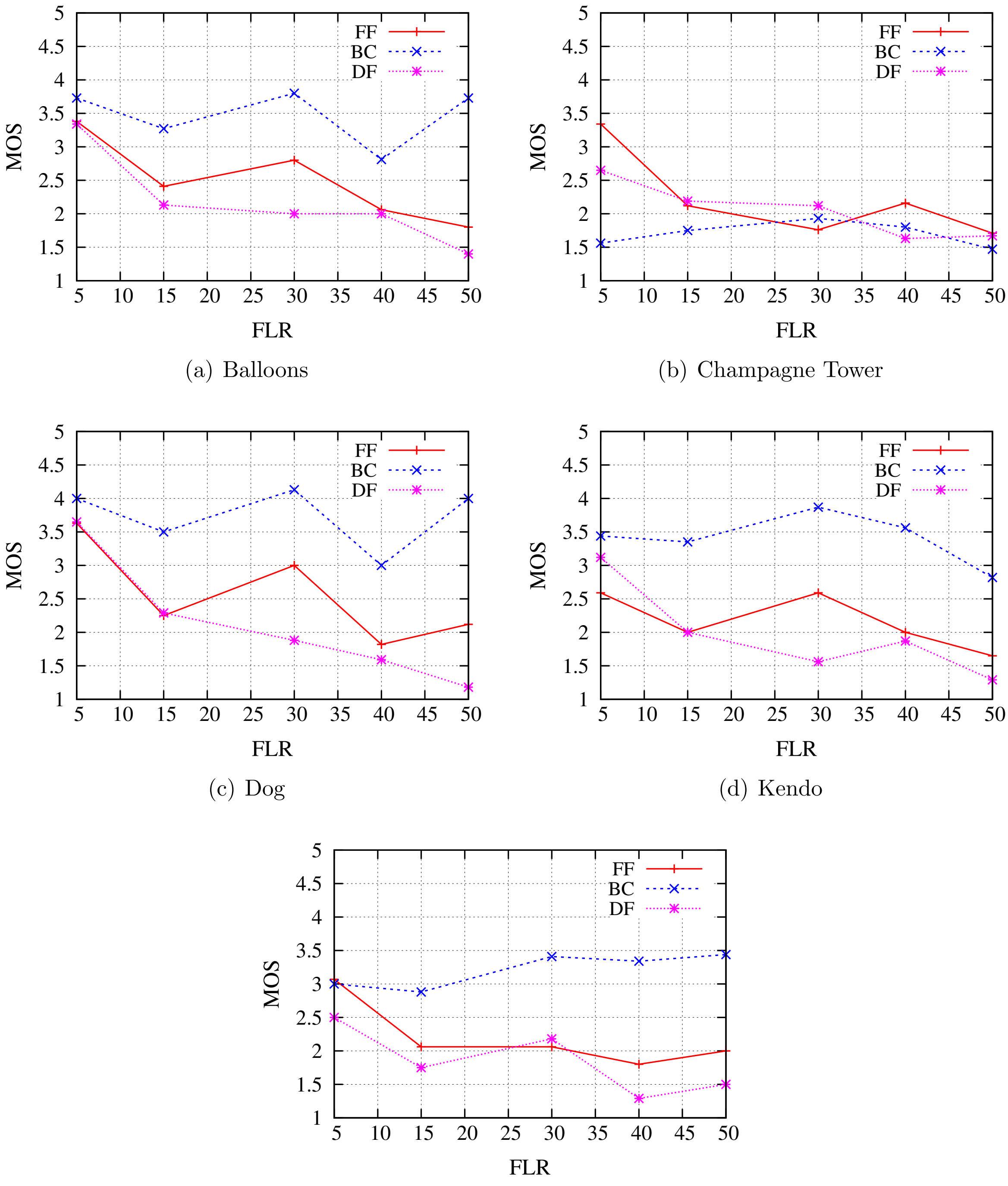
FLR (%)	N	Min	Max	Avg
46	11	2	22	7.6
38	10	1	14	6.9
29	7	1	33	6.7
18	4	4	15	8.0
6	2	5	6	5.5

generated for FLR=29% and the shorter (size=1) for both FLR=29% and FLR=38%. Taking into account the combinations resulting from the frame loss concealment methods and the frame loss probabilities, a total of 80 test sequences were generated and used in the subjective tests. The most relevant factors observed in the subjective study are discussed in the following Sub-sections.

Performance evaluation of the concealment methods

In this Sub-section the results regarding the viewers' MOS versus the FLR are analysed. Figure 4.11 shows theses results for each of the five sequences used in the experiments. In this figure, one may notice that, for very small error probabilities, all concealment methods present a similar MOS. However, as the number of losses increases (*i.e.*, higher FLR), the BC method rapidly presents the best MOS results, for all sequences, except for the highest disparity sequence, Champagne Tower. Note that the BC method tends to compromise the stereo effect by switching to 2D viewing during the error period, but maintains the motion information, by copying it from the base view. The results presented in Figure 4.11 also demonstrate that, for low to moderate disparity sequences, the users prefer to lose 3D perception than to endure motion related artifacts, like frame freezing.

One may also notice that, in this set of experiments, users tend to give higher scores to those sequences that present 30% of missing frames (FLR=29% in Table 4.2) . This can be explained by looking carefully at the error pattern generated in this case. These particular sequences were subjected to longer loss bursts, which resulted not only in longer error concealment periods but also in longer periods without losses. This factor benefits the BC concealment method, because it causes less transitions from 3D to 2D and vice-versa. However, one may also argue that it has a beneficial effect for the other methods, because it extends the periods with no errors. Nevertheless, further studies, with longer sequences, may be required to conclude about this particular factor, *i.e.*, how the random impairment length period, jointly considered with error concealment methods, influence 3D perceptual quality.



(a) Balloons

(b) Champagne Tower

(c) Dog

(d) Kendo

(e) Pantonime

Figure 4.11: MOS versus FLR for each concealment method and sequence.

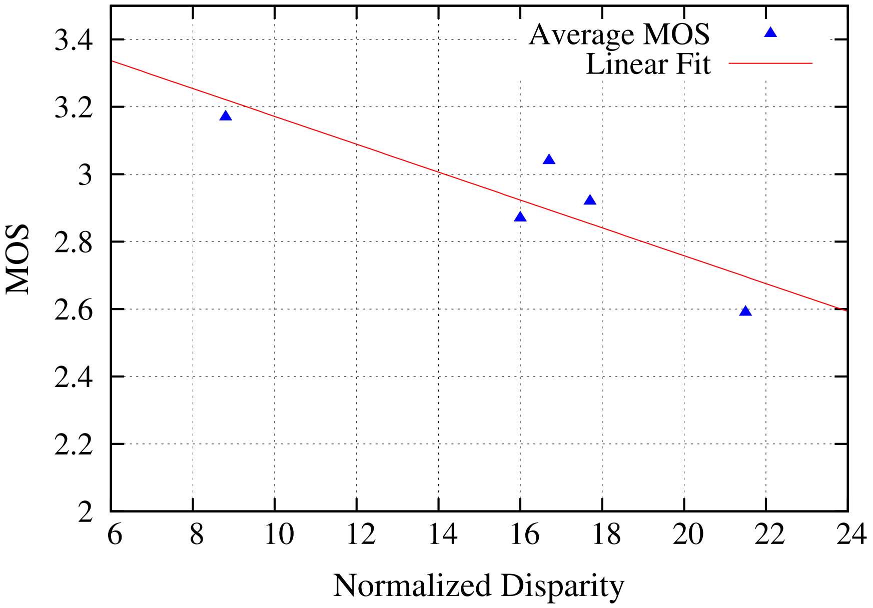


Figure 4.12: MOS versus sequence Disparity for random loss.

Disparity influence on perceived 3D experience

The general viewers' opinion can also be related with other factors. One of the relevant factors in 3D video is the video disparity, as previously found in [123, 124]. Figure 4.12 represents the average MOS, for all concealment methods, as a function of the disparity of the corresponding test sequence (as given in Table 4.1).

As for the case of regularly spaced frame loss, in Figure 4.12, one may observe that, in the presence of random errors, viewers tend to change their initial evaluation (*i.e.*, higher scores are given to higher disparity sequences) and grade low disparity sequences with higher values of MOS. This means that viewers tolerate better 3D errors that have smaller stereoscopic visual impact. These results show that, for high disparity sequences, the stereoscopic errors are more noticeable.

Figure 4.13 presents a more detailed analysis. The MOS is represented as a function of the disparity of each sequence, for each of the concealment methods implemented in this study. Again, one may observe that, on average, the BC concealment method is able to outperform the FF and DF schemes, for low to medium disparity sequences. Nevertheless, one may also notice that, for high disparity values, the MOS of the sequence processed with BC decays severely. This is because users tend to give higher significance to errors in sequences where the 3D effect (*i.e.*, depth perception) is subjectively higher. In other words, as pointed out in the previous case of regular frame loss, losing 3D perception in sequences with low perceived depth is less annoying than in sequences with high perceived depth.

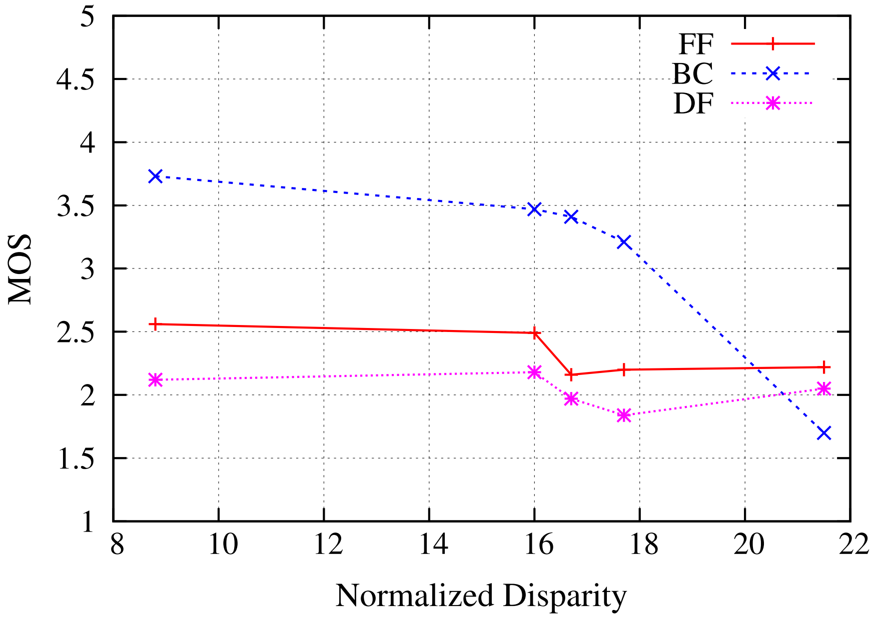


Figure 4.13: MOS versus sequence disparity and frame loss concealment method.

Figure 4.14 presents the MOS obtained for each concealment method, for all the test sequences, as a function of the FLR. One may notice a clear advantage of the BC error concealment scheme, for all error probabilities values. Figure 4.14 also shows that, for increasing values of FLR, the results of the FF and DF methods tend to decay. On the contrary, the MOS for BC concealed sequences tends to maintain a steady value. This fact further demonstrates the advantage of using BC for error concealment in the tested scenarios.

One should notice that the concealment methods used in this study are simple schemes that replace an entire frame for each missing frame at the receiver. The absolute MOS obtained with more efficient frame concealment methods (*e.g.*, using spatio-temporal and interview information) are expected to show better results than those obtained in this study. However, their relative behaviour and the main factors presented in this research work should remain valid for other concealment methods, because the perceptual experience is the same, though its effect might be smoothed using more complex concealment methods. Therefore these experimental results can be considered the lowest ground truth measure of the expected performance of more sophisticated receivers.

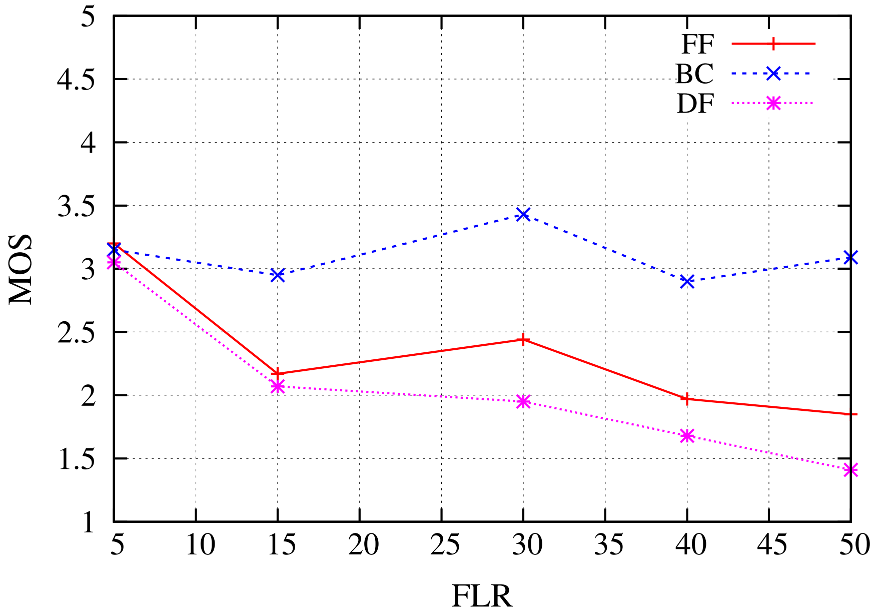


Figure 4.14: MOS versus FLR and concealment method.

4.5 Summary

This chapter presented a subjective evaluation study about the impact of frame loss in 3D video quality, considering sequence's disparity and concealment methods as relevant factors. The results clearly show that disparity can be used as a quality factor in 3D video. It was found that for low disparity sequences and short frame loss burst duration, just repeating the previously decoded frame in the auxiliary view is a good concealment method. However, for longer error bursts and higher disparity video, it is better to switch into 2D visualisation rather than trying to restore the 3D effect by using other frame concealment methods. Overall, these results also provide relevant insight for implementation of frame skipping decisions in 3D video rate control and perceptually efficient frame loss concealment strategies.

Chapter 5

Quality evaluation of 3DTV over hierarchical DVB-T

In the near future, the multi-view coding extension of H.264/AVC [5] will allow even more flexibility in 3DTV services, ranging from IP to digital terrestrial broadcasting, without losing backward compatibility to current 2DTV systems. In the case of 3DTV, the multi-view video coding (MVC) format enables transmission of several different views of the same scene and provides higher levels of immersive user interaction than conventional 3D stereo video. However, full deployment of such type of service still faces the problem of backward compatibility in both decoding equipment and broadcast networking. Therefore, a relevant factor for the success of the future 3DTV broadcasting services will be the level of integration with existing technology in both the operator side and the user equipment.

In this context, the multi-view extension of H.264/AVC and the hierarchical mode of DVB-T provide the enabling framework for multi-view 3DTV services without requirements of deep changes in the available technology. In [125] a system framework for 3DTV is proposed, using the hierarchical DVB-T system to achieve unequal error protection for video and depth. The video, which contains more relevant information, is transmitted over the high priority channel, adding more resilience to errors in the transmission channel. Hierarchical transmission for scalable video was also proposed in [126] and the results have shown a clear advantage of the hierarchical mode compared to non-hierarchical.

The work present in [127], provides relevant information about the relationship between video quality and packet loss, assuming that only the auxiliary view was subject to errors. The conclusions of this research showed that removing the whole 3D viewing experience, *i.e.*, switch to 2D, is better than concealing the auxiliary view in an attempt to maintain 3D. These are relevant results, because they show that 3DTV service availability is not necessarily a binary ON-OFF function. In fact there is an intermediate operational region, where 3D may not be available due to poor channel conditions, but 2DTV is still fully operational.

In this chapter a simulation study of 3DTV broadcasting over a hierarchical DVB-T channel is presented. The aim is to find a range of channel conditions (packet loss), that allow the multi-view 3DTV service to be available, within acceptable distortion levels. Below the lowest threshold, *i.e.*, at excessive channel degradation conditions, unacceptable distortion is introduced in the 3D viewing experience leading to the unavailability of such service. In this case, switching from 3D to 2D viewing is seamlessly provided to users.

This chapter is organised as follows. Section 5.1 describes the multi-view 3DTV transmission system based on hierarchical DVB-T. Additionally, the developed tools are presented, as well as a brief description of the system block to work with the proposed transmission approach. In Section 5.2 the experimental setup used to simulate the broadcasting system is presented. Moreover, three different quality model are devised from the analysis of the obtained results, showing good agreement with additional set of experiments. Finally, Section 5.3 concludes this chapter presenting a brief discussion of the results and the presented model.

5.1 System definition

In this Section we describe a solution for multi-view 3DTV broadcasting based on hierarchical DVB-T, as shown in the block diagram depicted in Figure 5.1. The base view is compliant with standard 2DTV systems using the H.264/AVC over MPEG-2 Transport Stream (TS) [52]. The remaining views are encoded using the standard extension H.264/MVC and multiplexed together into other TS. The 2DTV compliant TS is transmitted over the High-Priority (HP) channel of the hierarchical DVB-T modulator, while the multiple auxiliary views are transmitted over the Low-Priority (LP) channel. The

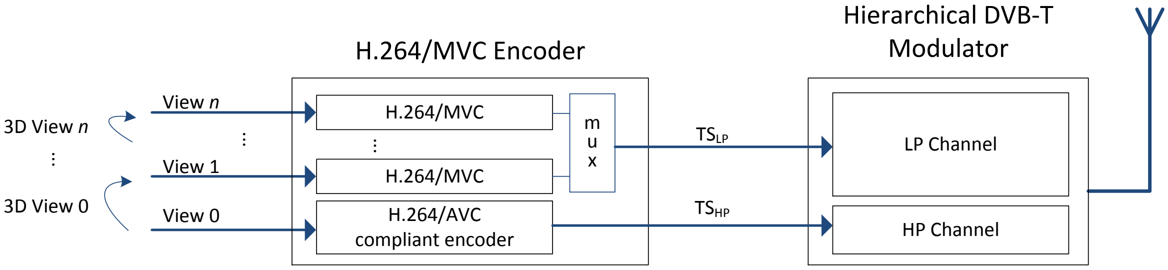


Figure 5.1: Block diagram of the hierarchical multi-view 3DTV transmitter.

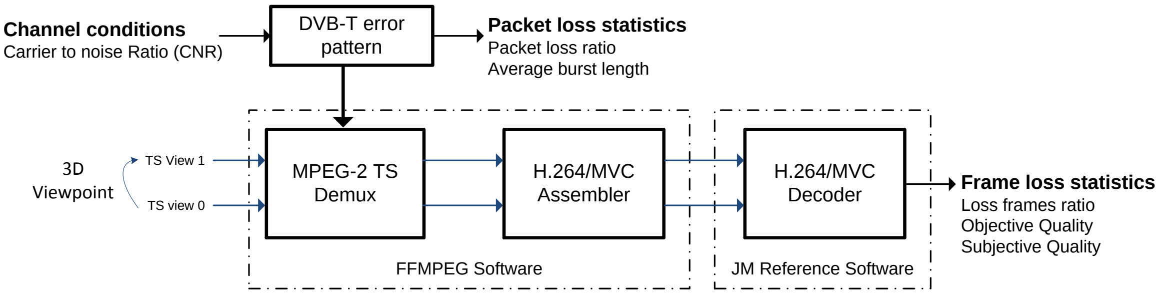


Figure 5.2: Block diagram of the implemented receiver for performance evaluation.

proposed approach has inherent backward compatibility with legacy 2DTV systems in the HP channel. The auxiliary views, sent through the LP channel, can be combined into stereo pairs at the decoder side, in order to provide a 3D viewing experience through multiple stereo views.

In order to complete this system, the current tools compliant with 2D video, *e.g.*, the MPEG-2 TS multiplexer present in the FFMPEG open source application, were extended to cope with the emerging MVC format. These modifications support the new coding units, *e.g.*, new NAL unit types and new parameters to identify each stream and associated view number. Therefore, firstly the MVC stream is divided in different sub-bitstreams, which are multiplexed using the modified FFMPEG adapted to the new format. The encapsulation of the base view does not suffer modifications, in order to maintain the backward compatibility.

Figure 5.2 shows the block diagram used to evaluate the performance of a 3DTV system using objective metrics. In this figure only the decoder side is represented, which includes the MPEG-2 demultiplexed and the MVC assembler implemented in the FFMPEG, and the H.264/MVC decoder. As mentioned before, the base and the auxiliary views are transmitted within different MPEG-2 streams. The MPEG-2 de-multiplexer processes each stream independently and then synchronises them through the decoded time stamp (DTS) values. To deal with packet loss, each Packetized elementary stream (PES) packet is marked if it contains some erroneous transport packet, and each marked PES packet is discarded. Since each PES packet carries one coded video frame, whenever a packet is lost, the whole frame is lost. After de-multiplexing two streams they are assembled to form a compatible H.264/MVC stream. The decoder processes the MVC stream, by performing the detection of the missing frames, using an extension for multi-view of the method proposed in [76]. Then the missing frames are concealed using the last received frame of the same view, designated as the Frame-Copy (FC) method.

5.2 3D video quality model for hierarchical DVB-T

In practical communication systems, the use of full-reference methods to evaluate the signal quality along the transmission chain is not feasible. Therefore, the approach followed in this section was to devise non-reference methods, based on quality models driven by a set of parameters derived from the transmitted stream itself. Such models are used to estimate the quality of the 3DTV service over hierarchical DVB-T.

Different objective quality metrics can be used to evaluate the quality of impaired video sequences. In our study, three objective metrics that are commonly used to evaluate multiview video quality were chosen, *i.e.*, the Peak Signal to Noise Ratio [128], the Structural Similarity Index Metric (SSIM) [102] and a stereo sense metric (SSM) metric [112]. The experimental results obtained with these metrics were fitted to create empirical models, aiming to find a correlation between the channel conditions and the video quality.

The experimental results presented in this section were obtained using a DVB-T channel model for the LP channel, to simulate the network environment. The H.264/MVC codec was based on the JM 17.2 reference software, and the MPEG-2 multiplexer was based on the FFMPEG implementation described before.

In our experiments one 3D sequence was used, resulting from the concatenation of five different 3D sequences with a resolution of 1024 × 768 pixels and a frame rate of 25 Hz. The sequences have distinct degrees of complexity and motion, as well as different levels of disparity, as have been described in Table 4.1 of Chapter 4, and the first frame of each sequence is included in the Appendix A. The sequences were concatenated in the following order: Balloons, Champagne Tower, Kendo, Pantonime and Dog. The combined sequence was encoded with the H.264/MVC encoder using five high-quality rates, a random access delay (IDR period) of half second and four reference frames. In Table 5.1 the quantisation parameters are presented, as well as the resulting rate-distortion relation of the transmitted streams. Since we wanted to have high quality in both views, the base and auxiliary views have similar objective quality (PSNR).

Table 5.1: Rate-Distortion relation of transmitted streams.

QP			Base view		Auxiliary view	
I	P	B	Bitrate	PSNR	Bitrate	PSNR
23	23	25	2539.10	43.38	2010.60	43.90
26	26	28	1615.00	42.10	1261.00	42.51
28	28	30	1247.40	41.19	960.88	41.53
33	33	35	697.59	38.52	510.58	38.63

Bitrate - kbits/s — PSNR - dB

5.2.1 Overview of the empirical models

The empirical models were devised by fitting the experimental data into different functions in order to reach a single expression that provides the video quality for different channel conditions and coding rates. The fitting was performed in the Curve Fitting Toolbox of MATLAB, using different expressions with the aim of finding the best relationship between the packet loss probability and the quality metric results. The models will be presented subsequently, and were obtained by fitting the initial data using the following equation:

$$y = \frac{p1 \times x + p2}{x + q1}, \tag{5.1}$$

where x is the packet loss ratio (PLR) and y the approximated value of the quality metric. For each quality metric, four equations as the one presented in the Equation 5.1 were obtained. This process generated four groups of coefficients $p1$, $p2$ and $q1$, one for each bitrate. The relation between the coefficients and the bitrate was evaluated and fitted, to reach a single expression, using the following exponential equation:

$$y = a \times x^b + c, \tag{5.2}$$

with x corresponding to the bitrate (R) and y corresponding to the values of the coefficients.

To create such empirical models, two sets of experiments were performed: one to obtain the experimental data and another to validate the models. Both experiments used the same channel conditions, but different random error patterns. The packet loss probabilities were chosen to cover the range of interest of the carrier-to-noise ratio, *i.e.*, from error free reception until loss of synchronism.

The next Sub-sections will present the results of the studied quality metrics and the obtained models using the previous methodology.

5.2.2 Peak Signal to Noise Ratio (PSNR) model

Peak Signal-to-Noise Ratio (PSNR) is a quality metric that is commonly used to measure the quality of 2D videos. Besides that, the PSNR is also used to evaluate the 3D video quality [5] [18]. Thus, the temporal distortion of the received 3D video stream transmitted over the hierarchical DVB-T channel is evaluated with the PSNR metric, in order to estimate an empirical model to predict the distortion based on the channel conditions.

Figure 5.3 shows the PSNR obtained for the 3D video, as a function of the PLR. As observed in the figure, the temporal distortion of the received auxiliary view video increases with the packet loss ratio, and the 3D video streams with higher data rates

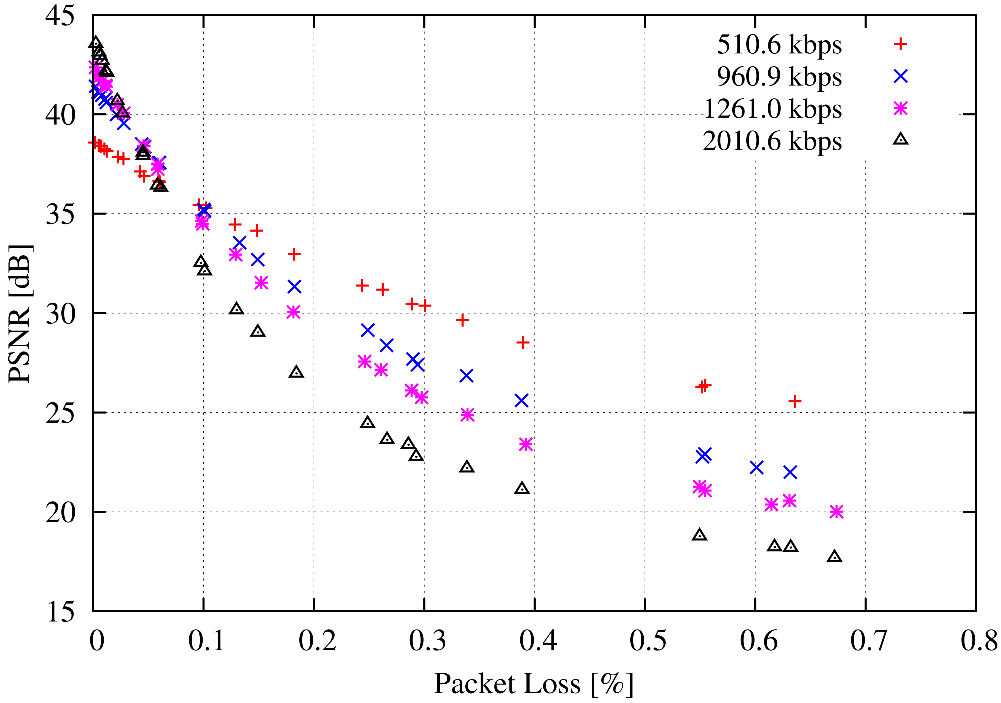


Figure 5.3: Temporal distortion evaluated using PSNR.

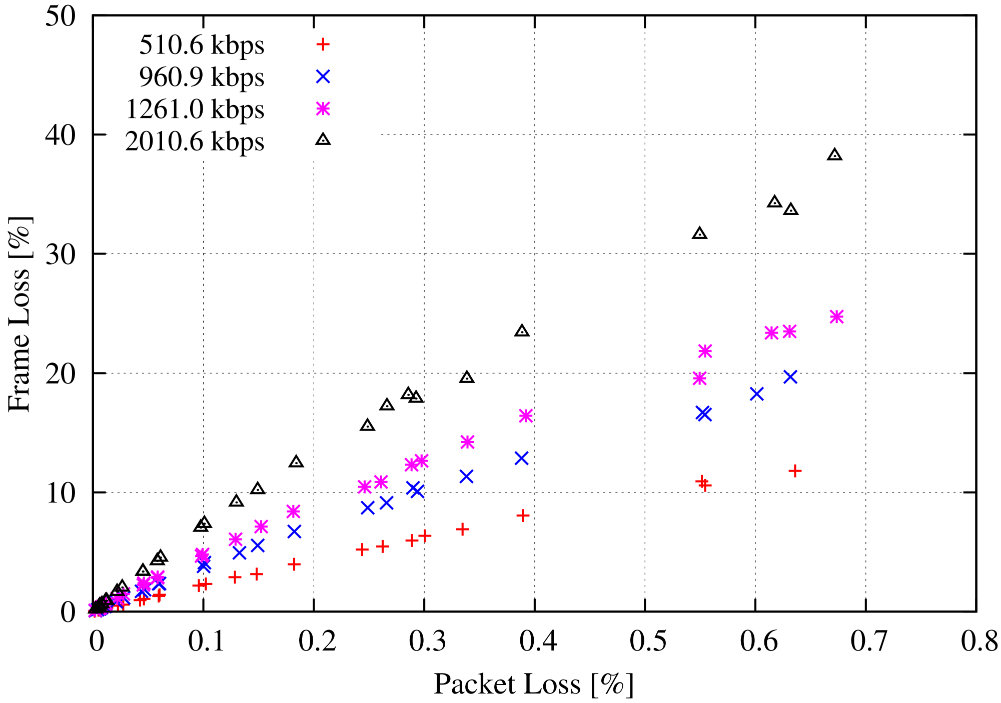


Figure 5.4: Results of frame loss ratio versus packet loss ratio.

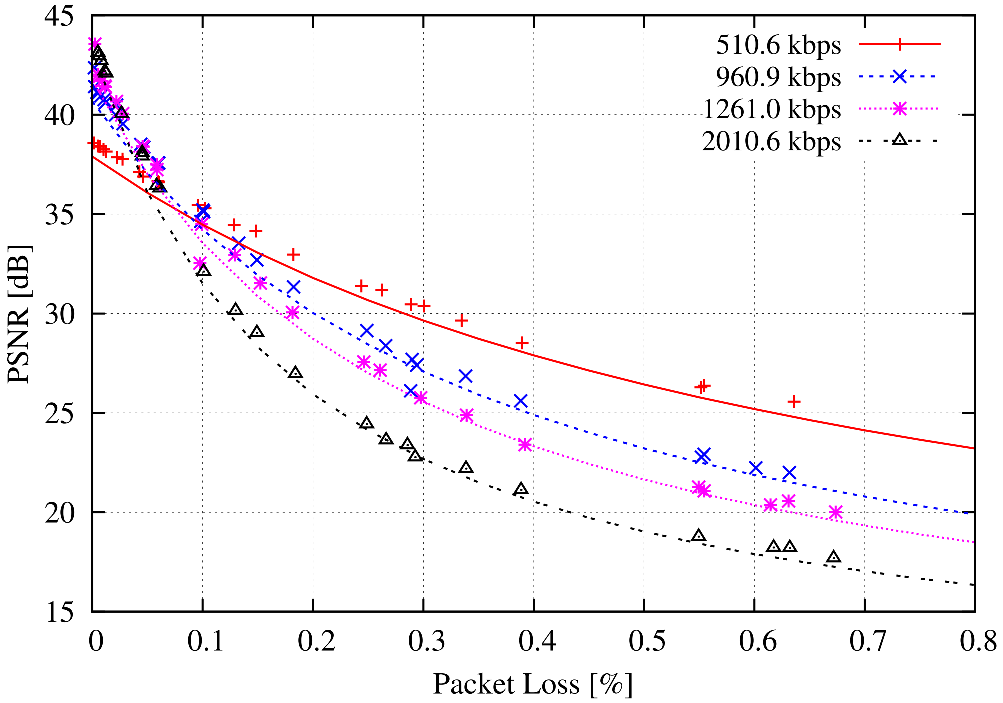


Figure 5.5: Empirical PSNR model (lines) and the experimental results (points).

are more affected by transmission errors, *i.e.*, lower PSNR. This is because, for the same PLR, more packets are likely to be affected by errors, as expressed in the results of Figure 5.4. The figure shows the number of lost frames for different packet loss ratios, and it can be seen by the results that for higher bitrates, more frames are affected by the transmission losses. In Figure 5.3 a different result occurs for lower loss probabilities, where the 3D videos with higher rates achieve higher quality, as expected, since they have higher error-free quality. This difference happens because increasing the bitrate leads to increasing of the length of PES packets, and longer packets are more prone to errors, which affects the video quality.

The analysis of the initial fitting of the PSNR results using the Equation 5.1 showed that the coefficient $p1$ did not have significant variations and did not show a relation with the bitrate. So, the average value of this coefficient was determined, in order to simplify the final expression. Then, for the coefficients $p2$ and $q1$ two expressions were obtained using Equation 5.2. This results in the final empirical model for the PSNR expressed as

$$PPSNR = \frac{10.22 \times PLR + 8175 \times R^{-0.92} + 0.39}{PLR + 508 \times R^{-1.06} + 0.022}, \tag{5.3}$$

where PPSNR represents the predicted PSNR value, which showed to be good match with the experimental data.

In Figure 5.5 a comparison between the proposed model and the experimental results is presented, where the lines represent the PSNR model and the dots represent the experimental data of the second set of experiments. The figure shows the strong correlation between the model and the obtained results. An objective analysis was also performed for the proposed model, showing an average root mean square error (RMSE) of 0.52 dB and a maximum absolute difference of 0.9 dB. These results indicate that the proposed model can predict the values for the PSNR of received 3D video over a DVB-T hierarchical channel.

5.2.3 Stereo sense metric (SSM) model

As shown by previous studies [112, 124, 129], there are objective metrics able to measure the observer's stereo perception of a given stereo video. The study in [124] showed that disparity information can be used as a quality metric, since it is related with depth perception. A combination of PSNR and disparity was also evaluated in [129], revealing good consistency with subjective results.

In [112] the relevance of the disparity (D) between the left (L) and right (R) images was evaluated, using a measure given by $D = \mid R - L \mid$. A constant noise was added to both left and right images, and the perceived quality was assessed. The results showed that when noise is added to regions with a high D, observers loose the stereo sense and notice a degradation in the image quality. When the noise is added to regions with a lower D, the observers just notice the image degradation, however they keep the stereo perception. Thus, the results demonstrate that the value D is related to depth perception (stereo sense). Based on this conclusions, a quality metric for stereo sense (SSM) is proposed, by measuring the distortion present in regions with higher disparity values. To measure the SSM the disparity (D) are firstly measured. Then, the slight noise from the disparity is removed, based on a magnitude threshold, *i.e.*, the points with magnitude lower than a given threshold are eliminated. The set of relevant points is represented with a matrix M. Finally, the SSM value is obtained using:

$$SSM = 10 \times log_{10} \frac{255^2}{MSE_M}, \tag{5.4}$$

with,

$$MSE_M = \frac{\sum_M \left[X_o(x, y) - X_r(x, y)\right]^2}{NoM}, \tag{5.5}$$

where X_o and X_r correspond to the original and reconstructed frames respectively, and NoM is the number of elements of the matrix M. The MSE_M represents the mean square error between the original and reconstructed images in the points with higher disparity values, marked by M.

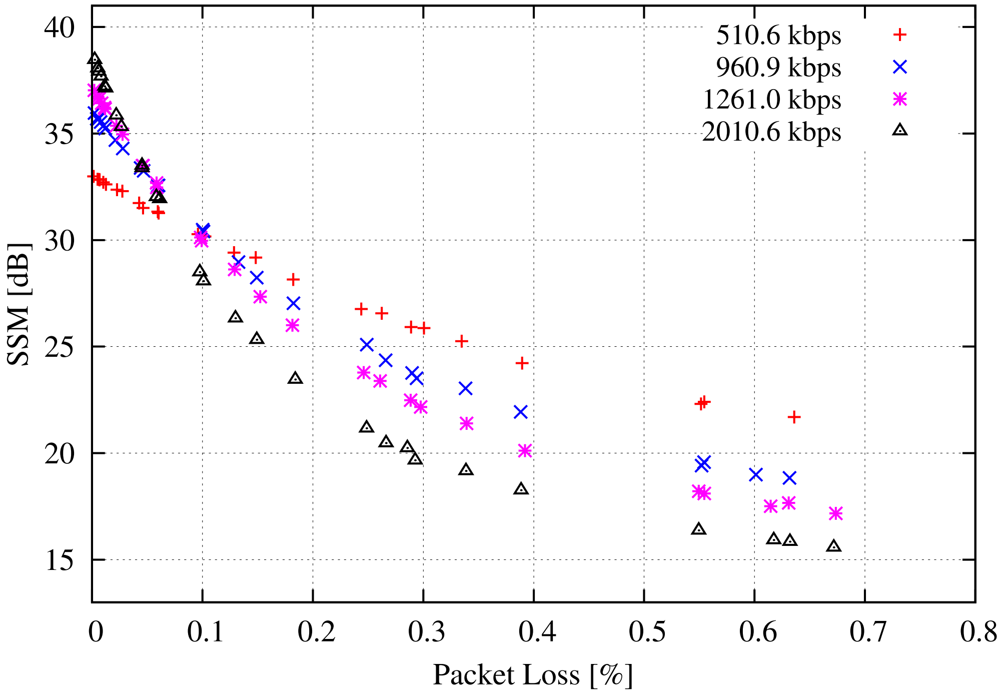


Figure 5.6: Results of temporal distortion of stereo perception.

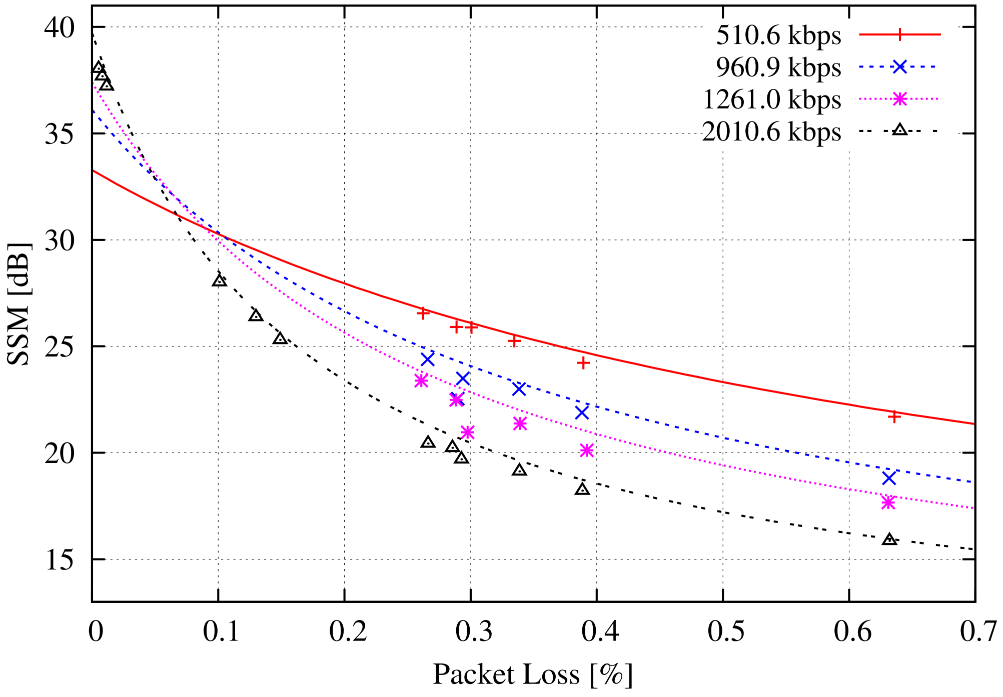


Figure 5.7: Comparison of the empirical model for the SSM (lines) with experimental results (points).

Figure 5.6 presents the results of SSM metric for different packet loss probabilities. This figure shows that the results of SSM metric are consistent with the PSNR (Figure 5.3), which indicates that the stereo perception decreases with the increasing PLR, as expected. One should notice that both PSNR and SSM are based on the MSE between the original and reconstructed images. However, the SSM metric focuses in the MSE of pixels with higher disparity, as they are more relevant to the stereo perception. Therefore, comparing the results of Figure 5.3 and Figure 5.6 reveals that stereo perception is more sensitive to coding and transmission errors, since SSM achieves lower values than PSNR.

The method used for PSNR results was applied to SSM results, creating a model that can predict the stereo quality for different amounts loss rates, using the following equation:

$$PSSM = \frac{9.58 \times PLR + 6096.0 \times R^{-0.895}}{PLR + 409.52 \times R^{-1.024}}, \tag{5.6}$$

where $PSSM$ is the predicted values of the SSM metric. Once again, this model shows a good matching with the experimental data, *e.g.*, a comparison with the second set of results show a maximum RMSE of 0.96 dB for the transmitted video stream with 1261.0 $kbits/s$, as can be seen in Figure 5.7.

5.2.4 Structural similarity index metric (SSIM) model

The SSIM metric, proposed by Zhou Wang [130], differs from the common objective quality metrics, by using the structural distortion measure instead of signal to noise ratio. This method is based on the idea that the human vision system is highly adapted in extracting the structural information from images [131]. Previous studies shown a good performance of SSIM [132] to evaluate the quality of 2D images. The usage of this metric for 3D video was also tested with acceptable results [133].

So far the results presented correspond to pixel distortion metrics, like PSNR. This Sub-section presents the results using structural distortion, that is the proposed model for the SSIM metric. As can be seen in Figure 5.8 the structural distortion (SSIM) decreases with the PLR growing, as occurred with previous metrics.

The empirical model for the SSIM metric prediction based on the Equations 5.1 and 5.2, is expressed as

$$PSSIM = \frac{P1_{SSIM} \times PLR + P2_{SSIM}}{PLR + Q1_{SSIM}}, \tag{5.7}$$

where the parameters $P1_{SSIM}$, $P2_{SSIM}$ and $P3_{SSIM}$ can be devised using:

$$P1_{SSIM} = -5.36 \times 10^{-17} \times R^{4.6} + 0.68, \tag{5.8}$$

$$P2_{SSIM} = 5.4 \times 10^{10} \times R^{-4.06} + 0.408, \tag{5.9}$$

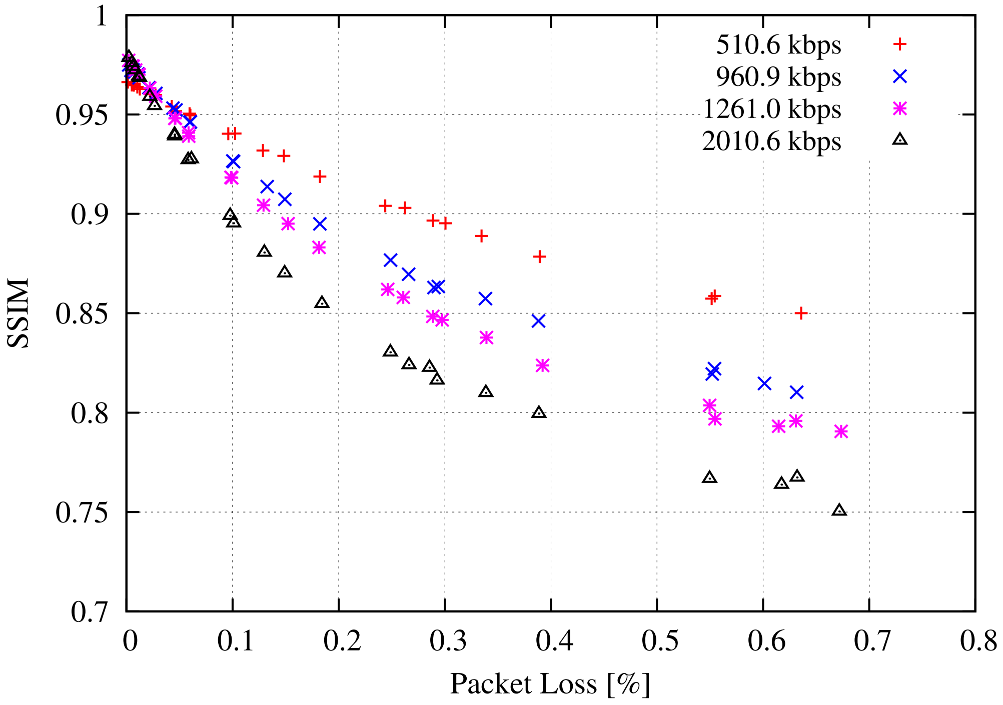


Figure 5.8: Structural distortion of the received auxiliary view stream.

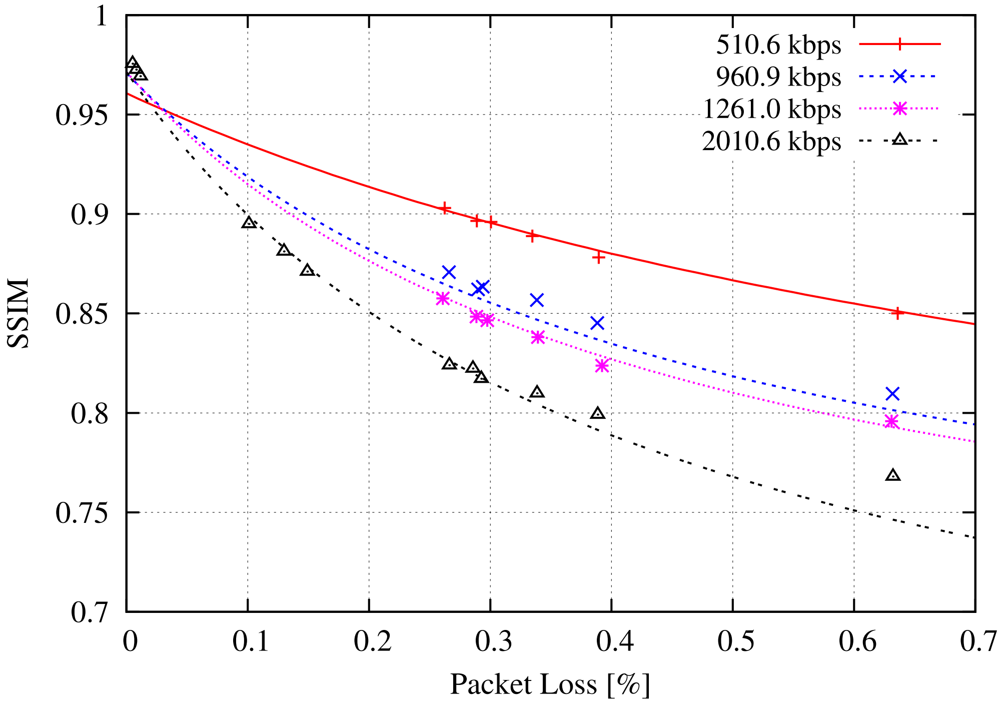


Figure 5.9: Comparison between the proposed model (lines) and experimental results (points).

$$Q1_{SSIM} = 6.42 \times 10^{10} \times R^{-4.08} + 0.42. \tag{5.10}$$

The matching between the proposed model (represented by lines) and the experimental data (represented by dots) can be observed in Figure 5.9. The objective analysis revealed an average RMSE over different bitrates of 0.00501. This result gives an indication about the model accuracy to predict the structural distortion when the reference information is not available.

5.3 Summary

In this chapter it has devised a set of experimental models, which can be applied to any video broadcast network using MPEG-2 transport stream over error prone channels. These models were devised using a simple frame loss concealment method, so they are able to measure the quality for the worst scenario, *i.e.*, lowest performance. Therefore, the models can be useful for the simplest and for the most sophisticated 3D video receivers compliant with the MPEG-2 transport stream and the H.264/MVC standard. Since these models show a good fidelity with the experimental data, being able to predict the impact of packet losses in the 3D video quality, they provide relevant insight to 3DTV network planning using the hierarchical DVB-T mode.

Finally, it can be concluded that the research work presented in this chapter leaded to an efficient 3D video receiver, enabling the objective quality assessment of 3D video affected by packet losses. Besides the complete development of the video receiver, extended research should be performed to find the relevant quality thresholds, where the 3D video quality is not acceptable and only the 2D video service is available. This would help to define the limits of the channel conditions that should be used to enable the 3D video services, when the hierarchical DVB-T modulation is used.

Chapter 6

Joint motion-disparity frame loss concealment method for 3D video

In generic multi-view video communications, the state-of-the-art codec H.264/MVC [5] is currently used in most stereoscopic video applications. Since this is a predictive coding format, MVC compressed streams are very sensitive to transmission errors and packet losses [13]. Whenever a frame is lost, error propagation through dependent frames contributes to increase the degradation of the reconstructed video quality and, consequently, the quality of experience.

The main goal of this chapter is to contribute with efficient techniques to cope with errors in the stereoscopic video coded using H.264/MVC, controlling the error propagation, during transmission. The concealment strategies presented in this chapter aim to recover from the loss of a whole frame, not only few blocks. Therefore, no neighbour information is present to be used as reference. Instead, the methods described exploit the correctly received information in the temporally close frames. There are also some basic monoscopic methods, such as frame copy, that can be applied to stereoscopic video in the case of frame loss. In order to improve the performance of the concealment, methods presented in [87, 89, 90], that use the assumption of constant motion, can also be applied. However, for coders that combine temporal and inter-view compensation, the temporal motion vector extrapolation methods may have a poor performance, since they cannot derive properly the inter-view vectors. Therefore, methods that utilise both temporal and inter-view information (vectors or an estimated disparity map) may increase their performance. The proposed method also aims to overcome the illumination problems of some recently proposed error concealment strategies, such as the methods presented in [93, 96, 97].

This chapter presents different concealment techniques devised in this research work. Section 6.1 describes three concealment schemes, which exploit the information presented in the error-free frames to recover the missing ones. The discussion of their performance for different sequences, under distinct error conditions, is presented in Section 6.2. Sections 6.3 and 6.4 proposes two methods resulting from the combination of the schemes presented in the Section 6.1. The comparison between both methods is also presented. Finally, a motion based decision approach is described in Section 6.5, and the Section 6.6 concludes the chapter.

6.1 Description of the concealment schemes

In this research work three concealment schemes were implemented and analysed. It is assumed that the base view is not corrupted, *e.g.*, by using a high priority transmission channel as described in Chapter 5. The different concealment schemes were used to reconstruct the missing frames in the auxiliary view of the 3D video, compressed with the H.264/MVC standard.

Figure 6.1 illustrates the information used by the different concealment schemes implemented and tested in the decoder. These schemes use two stereoscopic views, *i.e.*, the base view frames (f^0) and the auxiliary view frames (f^1), the disparity map (D), the motion vectors (mv) and disparity vectors (dv). In the figure, the lost frame is represented with a dashed square.

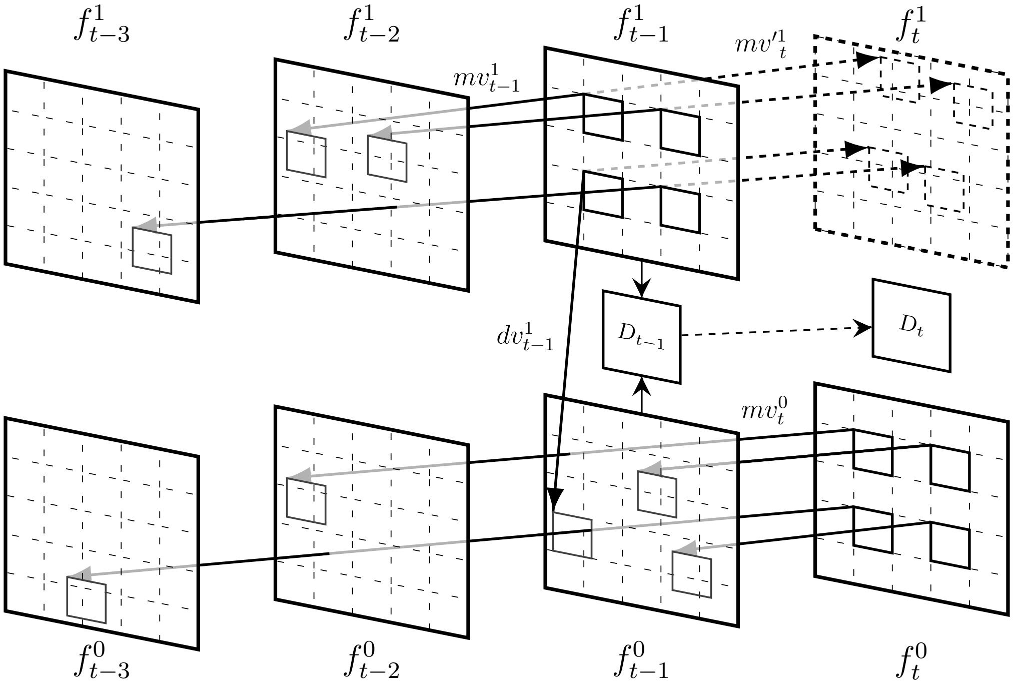


Figure 6.1: Representation of the base and auxiliary view frame, the corresponding motion/disparity vectors and the disparity map between the two stereoscopic views.

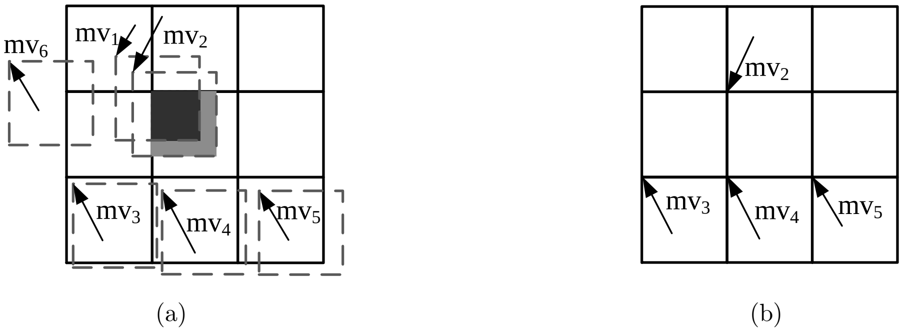


Figure 6.2: Original position of the vectors after projection (a); remaining vectors, associated with the 4 × 4 grid of the picture (b).

6.1.1 Extrapolation of the available vectors in the auxiliary view

The first concealment scheme, referred to as MVE, extrapolates each motion vector (MV), mv^1_{t-1}, from f^1_{t-1} onto f^1_t, in order to estimate a new set of MVs, mv'^1_t, for the missing frame. Note that the MVs of f^1_{t-1} may point to different reference frames (*e.g.*, f^1_{t-3} and f^1_{t-2} in Figure 6.1). The extrapolated MVs, mv'^1_t, are determined as follows:

$$mv'^1_t(x,y) = -\frac{t - t_C}{t_C - t_R} \times mv^1_{t-1}(x,y), \tag{6.1}$$

where t_R is the time instant of the reference frame pointed by $mv^1_{t-1}(x,y)$, and t_C refers to the time instant of the frame from which the MVs are extrapolated (*i.e.*, $t-1$). The dotted squares in the lost frame f^1_t, shown in Figure 6.1, represent the blocks of f^1_{t-1} projected by each mv'. Then, each extrapolated MV is associated with the closest 4 × 4 block of f^1_t that matches a fixed grid (also shown in Figure 6.1).

As a result of the previous process, several motion/disparity vectors may be associated with a given block. Since only one vector can be used for each block, the chosen one is determined based on the size of the overlapped area between the projected block and the final block that matches the fixed grid, *i.e.*, the vector associated with the larger overlapped area. Figure 6.2(a) represents several different cases of extrapolated MV and their correspondence to the fixed 4 × 4 grid. For example, mv_1 and mv_2 are associated with the central block of the represented grid, since it is the closest one. In this case, mv_2 is selected as the best MV because its projected block has a larger overlapped area (represented in light grey) than the projected block pointed by the mv_1 (represented in dark grey). The MVs, mv_3, mv_4 and mv_5, are chosen since they are the only vectors associated with the corresponding blocks on grid. Moreover, if an extrapolated MV is associated with a region that falls outside the image, it is discarded (*e.g.*, mv_6). The final MVs and their associated blocks on the grid are represented in Figure 6.2(b).

A particular case is employed when the frame f_{t-1}^{1} use disparity compensation for a certain block, rather than motion compensation (*i.e.*, the compensation vector points to f_{t-1}^{0}). Those blocks have a disparity vector (dv_{t-1}^{1}) associated to them, instead of a MV. Therefore, a dependency between those vectors and time does not exist to perform the temporal extrapolation. The disparity vector (DV) represents a shift of the auxiliary frame's pixels, when compared with the corresponding one in the base view. In this work the disparity vectors are assumed constant in a short time interval, as previously assumed for motion. Thus, each DVs present in f_{t-1}^{1} results in a new one at the missing frame ($dv_{t}'^{1}$), that will be used to fetch the pixels from the corresponding frame in the base view, f_{t}^{0}.

Finally, one should note that this process does not ensure MVs for all blocks. Consequently, some of them may end up without any associated MV. For those blocks, the average of the three top-left neighbours' vectors is used.

6.1.2 Disparity compensation of the base view MVs

The second concealment scheme uses the MVs of the base view frame to estimate the motion field of the corresponding lost frame in the auxiliary view. This concealment approach is referred to as BASE_MC. As shown in Figure 6.3, the motion field for the lost frame in the auxiliary at the instant t (f_{t}^{1}) is recovered using a disparity compensated version of the MVs (mv_{t}^{0}) of the co-located frame in the base view (f_{t}^{1}). Then, the reconstructed motion field is used to recover the missing frame from the preceding frames of the auxiliary view (*i.e.*, f_{t-1}^{1} and f_{t-2}^{1}).

This process requires disparity compensation, in order to determine the MVs for the auxiliary view from those of the base view. In this section the disparity estimation is first described. Finally, the process to fetch the MVs from the base view are presented.

Disparity Estimation

The disparity map for the last error-free stereo pair (D_{t-1} in Figure 6.1) is first obtained by using the variational method described in [134]. The method uses the Euler-Lagrange equation to minimise the energy function given by:

$$E(u(x,y)) = \int\int_{\Omega} \|I_1(x,y) - I_2(x+u(x,y),y)\|^2 + \varphi \times \psi\left(|\nabla u(x,y)|^2\right) dS, \qquad (6.2)$$

where I_1 and I_2 are the stereo image pair with rectangular image domain Ω. $u(x,y)$ corresponds to the disparity of each pixel (x,y) and φ is a weighting factor for the smoothness term $\psi\left(|\nabla u(x,y)|^2\right)$ [92]. The smoothness term is derived from the assumption that the

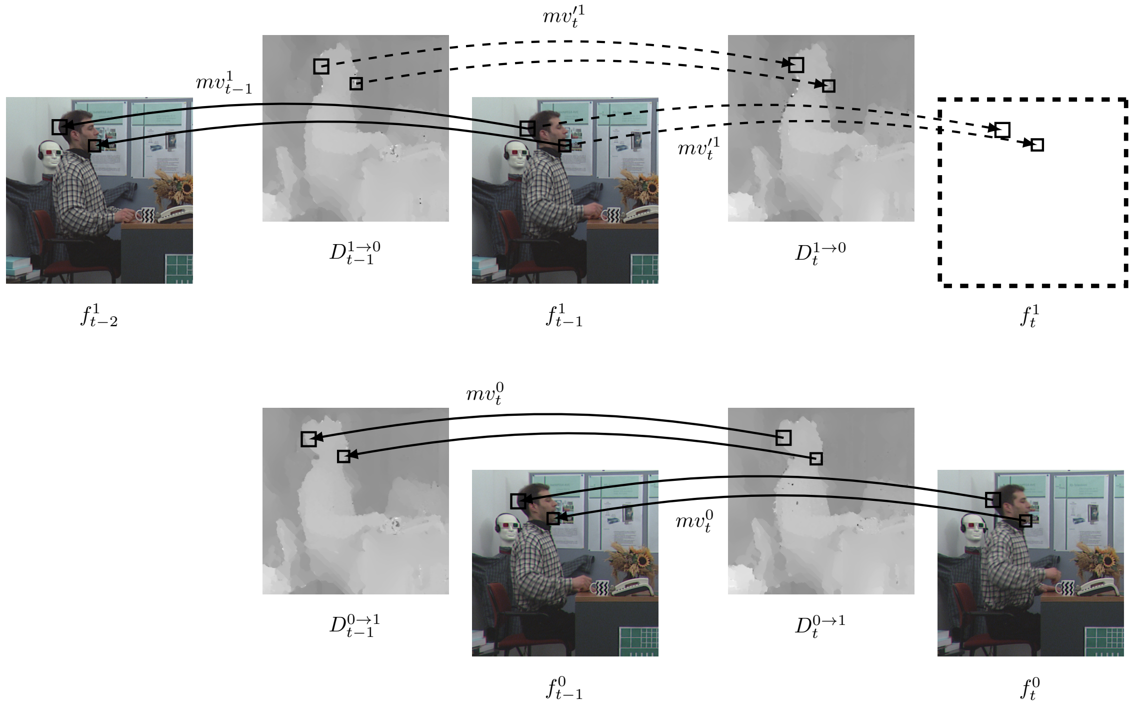


Figure 6.3: Representation of a stereoscopic sequence with associated disparity map and corresponding available MVs

neighbouring regions belong to the same object and thus these regions have similar disparities. This method has been used due to its good performance, comparing with block matching algorithms.

One should notice that the disparity map can be estimated using one of two possibilities. On the one hand, the pixels in the auxiliary view are searched in the base view frame. Thus, the disparity map is associated with the auxiliary view, and is referred to as $D^{1\to0}$. In this case the following applies:

$$f_t^1(y,x) = f_t^0(y, x + D_t^{1\to0}(y,x)). \tag{6.3}$$

On the other hand, the disparity map is obtained by searching the pixels from the base view frame in the co-located frame of auxiliary view. In this approach the disparity map is associated with the base view, and the following applies:

$$f_t^0(y,x) = f_t^1(y, x + D_t^{0\to1}(y,x)), \tag{6.4}$$

where $D_t^{0\to1}$ is the disparity map obtained in this case.

Figure 6.3 shows an example of a stereoscopic sequence, *i.e.*, the base and auxiliary view frames (f^0 and f^1) are illustrated, and the different disparity maps used in this research work. In the figure, $D_{t-1}^{0\to1}$ and $D_{t-1}^{1\to0}$ were obtained with the method in [134].

However, since f_t^1 is missing, an estimation of the disparity map for the missing frame, D_t, is obtained from a motion compensated version of D_{t-1}. On the one hand, if the disparity map is associated with the auxiliary view frame ($D^{1\to 0}$), the motion compensation uses the extrapolated MVs, $mv_t'^1$, as described in Sub-section 6.1.1 by the Equation 6.1. As shown in Figure 6.3, the temporally extrapolated MVs of the frame f_{t-1}^1 are used to project the disparity map from $t-1$ onto t. On the other hand, if the disparity map is associated with the base view frame ($D^{0\to 1}$), it is compensated using the base view vectors. As shown in Figure 6.3, the available vectors (mv_t^0) of the co-located frame in the base view (f_t^0) are used to obtain the disparity at instant t. In both of the cases, for those blocks without an associated MV, a zero MV is used, *i.e.*, the disparity value is copied from the co-located block of the disparity map of the previous time instant.

The estimated disparity map for the lost frame represents the pixels displacement between co-located frames of different views. Therefore, the block wise disparity map is determined by averaging the disparity values within the block area (*i.e.*, 4×4 disparity values).

Disparity compensation of the base-view motion field

As mentioned before, in this concealment scheme the motion field for the lost frame is obtained from the disparity compensated version of the base view MVs. Whenever possible (*i.e.*, MVs exist in the base view), the motion field for f_t^1 is determined by using the disparity compensated MVs obtained from the co-located base view frame, f_t^0. In order to accomplish this, one of the two following procedures is performed:

1. The vectors for the auxiliary view are fetched from the base view frame from a single block. Thus, the value of the disparity is rounded to match the 4×4 grid in the base view frame. Then, the following equation is applied:

   $$mv_t^1(x, y) = mv_t^0(x - D_t^{1\to 0}(x, y), y), \tag{6.5}$$

   where $D_t^{1\to 0}$ is the disparity associated with the auxiliary view frame.

2. The vectors of the base view are shifted according to the corresponding disparity values onto the auxiliary view frame, followed by a grid alignment (see Figure 6.2), to match the 4×4 fixed grid of the auxiliary view frame. This process is described as follows:

   $$mv_t^1(x - D_t^{0\to 1}, y) = mv_t^0(x, y), \tag{6.6}$$

   where $D_t^{0\to 1}$ is the disparity associated with the base view frame co-located with the missing one.

Finally, as used in the MVE concealment scheme, the regions without an associated MV are filled using the average of the three top-left neighbours' vectors, achieving a complete motion field for the missing frame.

6.1.3 Disparity compensation of the co-located base view frame

The third method, named INTERVIEW_COPY, uses the base view frame to provide the pixel information to estimate the co-located lost frame in the auxiliary view. A block wise approach is implemented using the disparity map D_t to compensate the blocks of the base view frame onto the missing one.

The disparity map at instant t is obtained as described in Sub-section 6.1.2. Then, through motion compensation using the extrapolated motion vectors in the auxiliary view frame at the instant $t-1$ (MVs extrapolated using the Equation 6.1), the disparity map is projected onto the missing instant, as represented on top of Figure 6.3. The disparity map needs to be referred to the auxiliary view, otherwise it is no longer valid to fetch the base view pixels onto the missing frame in the auxiliary view. Although the usage of a disparity map associated with base view frame is possible, it may result in several overlapped areas and gaps in the reconstructed frame

The approach implemented in this research work overcomes the problem of the overlapped and un-filled regions, through the following condition:

$$f_t^1(x,y) = f_t^0(x - dv_t^1(x,y), y), \tag{6.7}$$

where the dv_t^1 is obtained from the average value of 4×4 disparity values of D_t at the corresponding block position. For those regions where the disparity vector points to outside of the co-located base view frame, the current method is discarded and the MVE method is used instead.

6.2 Experimental evaluation of the different schemes

6.2.1 Setup

In this section the experimental setup used to evaluate the performance of the concealment algorithms proposed in this chapter is described. The setup described below is used to assess the performance of the different concealment schemes previously explained, as well as the methods explained in further Sections 6.3 and 6.4.

In order to test the proposed methods five well-known stereoscopic test sequences, with different resolutions and distinct types of motion and texture complexity were chosen.

Table 6.1: Description of stereoscopic video sequences used to evaluate the performance of the different concealment methods.

Sequence	Resolution	Description	Views used
Akko & Kayo	640 × 480	High translational motion, with two moving persons; moderate texture details; moderate depth.	B:50 - A:49
Balloons	1024 × 768	Complex motion with moving camera; rotational movements in the background; low texture complexity.	B:3 - A:2
Book Arrival	1024 × 768	Moderate translational motion; some temporal instants present high translational motion with the entry of another person in the scene; moderate depth, with static objects with different depth values; moderate texture complexity.	B:9 - A:8
Kendo	1024 × 768	High translational motion with two moving persons; moderate depth; moderate texture, however with lots of white regions and several persons on the background.	B:3 - A:2
Champagne Tower	1280 × 960	Low motion; high depth; complex texture with transparent objects.	B:40 - A:39

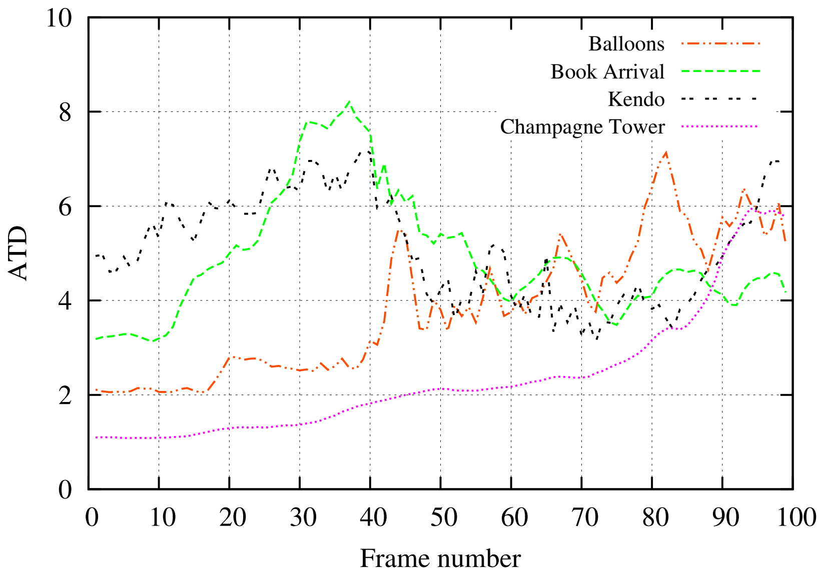


Figure 6.4: Temporal difference between consecutive frames for the test sequences.

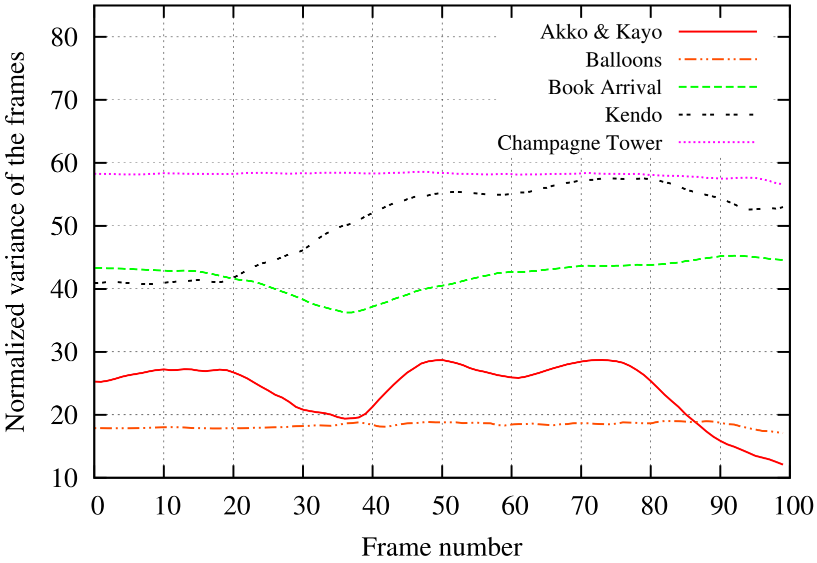


Figure 6.5: Spatial variance of each frame of the all test sequences.

These sequences also cover different kinds of disparity values. Table 6.1 presents a summary of the sequences' features, as well as the spatial resolution in pixels and the view IDs (*i.e.*, view number) used for the base (B) and auxiliary (A) view. The chosen test sequences were captured using 1D camera array. Since there is no rotation between cameras, the disparity map between two viewpoints can be correctly measured using the method implemented in [134]. The first frame of each sequence is illustrated in Appendix A.

The sequences' temporal complexity was evaluated using the absolute temporal difference *i.e.*, absolute difference of consecutive frames, for the first 100 frames of each sequence, which cover the frames used in these experiments. The ATD value for the frame at the instant t (f_t) is obtained as follows:

$$ATD(t) = \frac{\sum_y \sum_x |f_t(x,y) - f_{t-1}(x,y)|}{width \times height} \times 10^6. \tag{6.8}$$

Figure 6.4 shows the ATD results for the used test sequences, with exception of the sequence Akko & Kayo. This particular sequence has huge differences between consecutive frames (approximately 5 times higher than others), therefore, the results for the remaining sequences would be masked. The results show that sequences Kendo and Book Arrival (and Akko & Kayo) present higher motion than the other ones, especially than the sequence Champagne Tower.

In order to evaluate the texture's complexity of each sequence the pixels' variance within each frame is measured. In Figure 6.5 the complexity of the first 100 frames of each sequence used in the tests is presented. As illustrated in the figure, sequence Champagne Tower presents higher pixels' variance, in contrast with the motion, which may indicate that the concealment methods that depend on spatial redundancies may achieve lower performance than the ones that rely on temporal redundancies. In general the objective results for motion and texture complexity, presented in Figures 6.4 and 6.5, are correlated with the subjective evaluation presented in Table 6.1.

The test sequences were encoded using version 18.3 of the JM reference software [135], using Stereo High profile with an IDR period of 20 frames, a GOP structure IBPBP using 2 reference frames with inter-view prediction enabled, and a range of 64 pixels was used for motion/disparity search. The available coding modes in the H.264/MVC standard were used, and the same value for the quantisation parameter (QP) was configured for I-, P- and B-slices.

6.2.2 Performance evaluation

In this section the performance of the concealment methods is evaluated. The results show the objective quality (peak signal-to-noise ratio - PSNR), in dB, for the concealment of the missing frame at different coding rates (Mbits/s). The PSNR of the error-free frame is also presented, as a reference for comparison.

Firstly, the performance of the BASE_MC method using the two different approaches, represented by Equations 6.5 and 6.6, is analysed. This comparison intends to find the more suitable reference for the disparity map, *i.e*, base or auxiliary view. Table 6.2 presents the difference between the quality of the reconstructed frames using the base and auxiliary view MVs as reference to measure the disparity map. The quality of the concealed frame was evaluated using the PSNR (dB). The fourth column of the table (BASE_MC Base versus Auxiliary view) shows the quality differences between using the disparity map associated with the base view frame ($D^{0\to1}$) and associated with the auxiliary view frame ($D^{1\to0}$). One should notice that, if BASE_MC uses ($D^{0\to1}$), the MVs of the base view frame are used to temporally compensate the disparity map. In contrast, if ($D^{1\to0}$) is used, the temporal compensation uses the MVs of the auxiliary view frame (extrapolated MVs). A special case is also presented in third collum of the table (BASE_MC_REF Base versus Auxiliary view), where instead of temporally compensated the disparity map, it is obtained at the missing instant. This is performed through the stereo pair of the original sequence.

Results present in the third column of Table 6.2 show that both approaches, *i.e.*, using the disparity map associated with the base and auxiliary views, achieve similar performance for all test sequences. Note that, in this case the disparity map is obtained at the missing instant. However, a slight improvement is noticeable when the disparity map is associated with the auxiliary view frame ($D^{1\to0}$). This case uses Equation 6.5 to fetch the base view motion vectors, which achieves higher performance than using

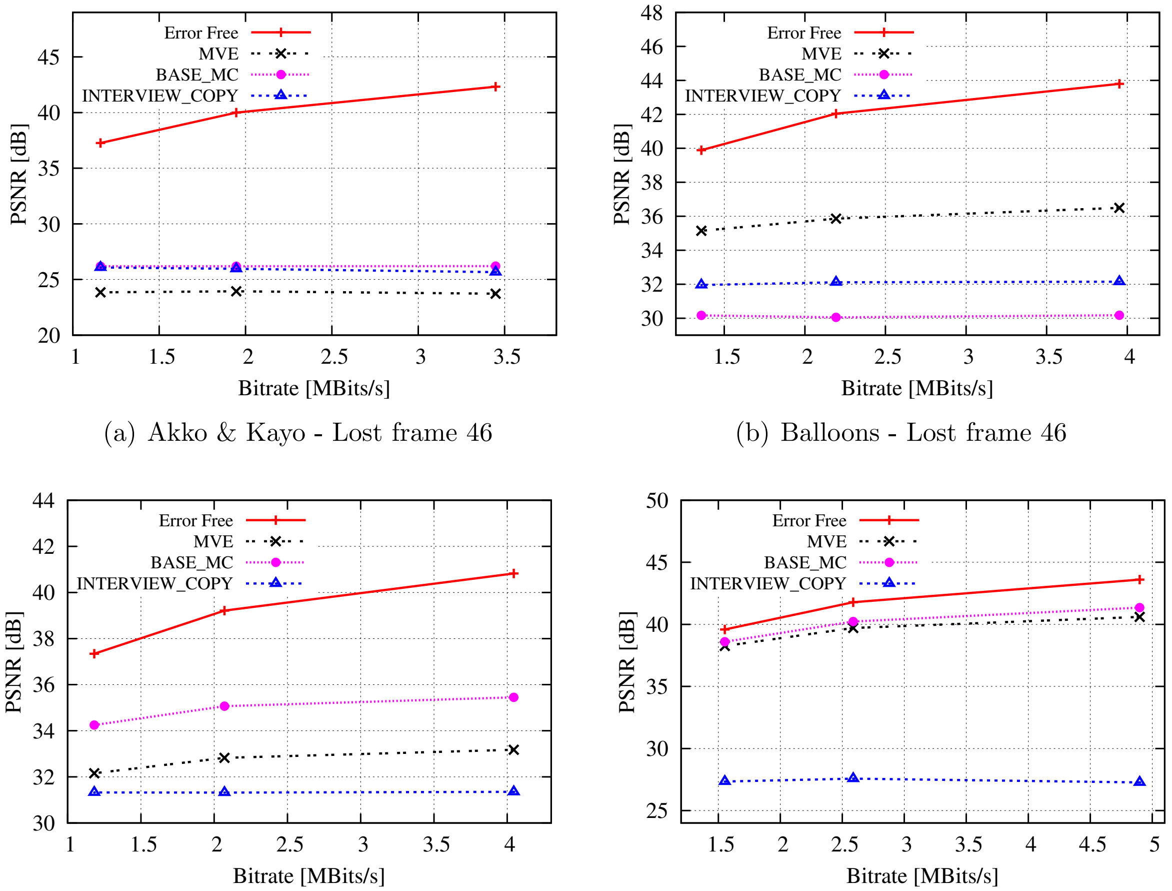
Table 6.2: Comparison of the reconstructed frame's quality (evaluated with the PSNR) when frame 6 is lost for BASE_MC concealment scheme using the disparity map associated with the base and auxiliary views.

Sequence	QP	BASE_MC_Ref Base versus Auxiliary view [Δ dB]	BASE_MC Base versus Auxiliary view [Δ dB]
Akko & Kayo	24	-0.444	0.810
	28	-0.453	0.437
	32	-0.425	0.806
Balloons	24	0.025	0.003
	28	0.059	0.151
	32	0.001	-0.055
Book Arrival	24	-0.124	-0.110
	28	-0.036	0.049
	32	-0.001	0.016
Champagne Tower	24	-0.352	0.04
	28	-0.115	0.069
	32	-0.121	0.201
Kendo	24	-0.323	-0.022
	28	-0.773	-0.501
	32	-1.149	0.062

Equation 6.6 with base view frame as reference for disparity estimation ($D^{0\to1}$). This indicates that Equation 6.5 is more suitable to reconstruct the missing motion field from the MVs of the co-located base view frame.

As mentioned before, Table 6.2 also shows the comparison of two possible approaches of the BASE_MC method using the temporally compensated disparity map (fourth column). In this case, the performance of the BASE_MC method using the auxiliary view as reference for disparity estimation decreases comparing with the opposite case. The comparison of the third and fourth column of Table 6.2 reveals that the difference increases for sequences with higher motion (Akko & Kayo and Kendo), indicating that the projection of the disparity map D_{t-1} onto the missing instant (D_t) using the base view vectors leads to higher quality in the reconstructed frame when the motion increases. This indicates that the base view motion vectors (mv_t^0) can project more accurately the disparity map, and, consequently, increase the performance of the BASE_MC method.

Secondly, all the techniques described in Section 6.1 are compared. The BASE_MC method in these experiments uses the disparity map projected using the temporally extrapolated MVs of the auxiliary view ($mv_t'^1$). Therefore, a fair comparison with the INTERVIEW_COPY method can be performed, since it relies on the same approach. In different frame loss events presented, the motion/disparity field is recovered using one of these three methods.



(a) Akko & Kayo - Lost frame 46

(b) Balloons - Lost frame 46

(c) Book Arrival - Lost frame 92

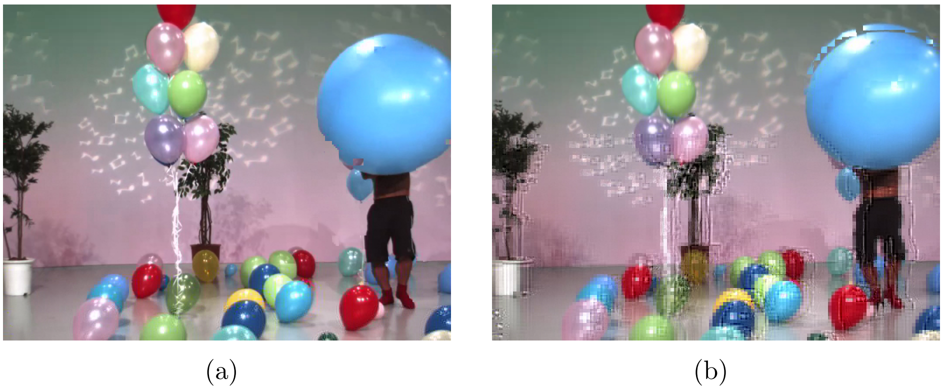
(d) Champagne Tower - Lost frame 6

Figure 6.6: Performance evaluation of the different concealment schemes when a P-Frame (reference frame) is lost.

Results presented in Figure 6.6 show that the reconstruction using the BASE_MC method achieves higher quality than the MVE method for most of the studied cases. This indicates that the base view vectors provide more accuracy than the ones in the auxiliary view. A particular case in the sequence Balloons when the frame 46 is missing, which is not cover in the figure, is further analysed. Table 6.3 presents the results for the two stages of the concealment methods. Stage 1 correspond to the usage of MVs provided by the concealment method (do not include the top-left MVs average), and the stage 2 corresponds to the final concealed frame. Results presented in Table 6.3 show that the BASE_MC method can achieve more 3 dB in the first stage. However, after the stage 2 the MVE method outperforms the BASE_MC method, indicating that the average of the three (top-left) neighbour blocks is not efficient in recovering the lost motion field. Figure 6.7 presents the reconstructed frame using these two concealment methods. The subjective analysis of the reconstructed frames reveals the clear advantage of the MVE method when compared with the BASE_MC. On the top of the big blue balloon in Figure 6.7(b), several

Table 6.3: Comparison of the methods MVE and BASE_MC for the sequence Balloons when the frame 46 is lost.

Stage	MVE		BASE_MC	
	Pixels [%]	PSNR [dB]	Pixels [%]	PSNR [dB]
Only MVs from the concealment method (Stage 1)	86.38	35.91	78.82	37.59
Using Average MVs (Stage 2)	100	35.995	100	30.07



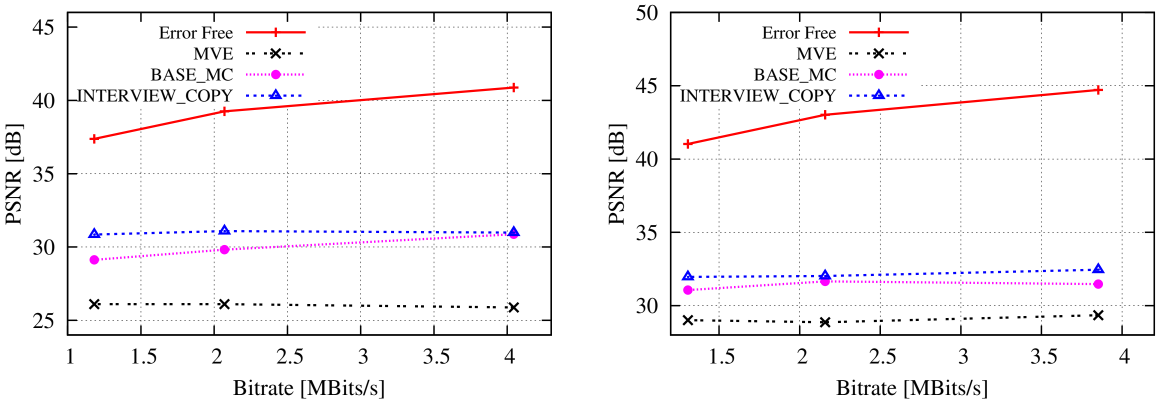
(a)

(b)

Figure 6.7: Subjective results of the reconstructed frame 46 for the method MVE (a) and BASE_MC (b).

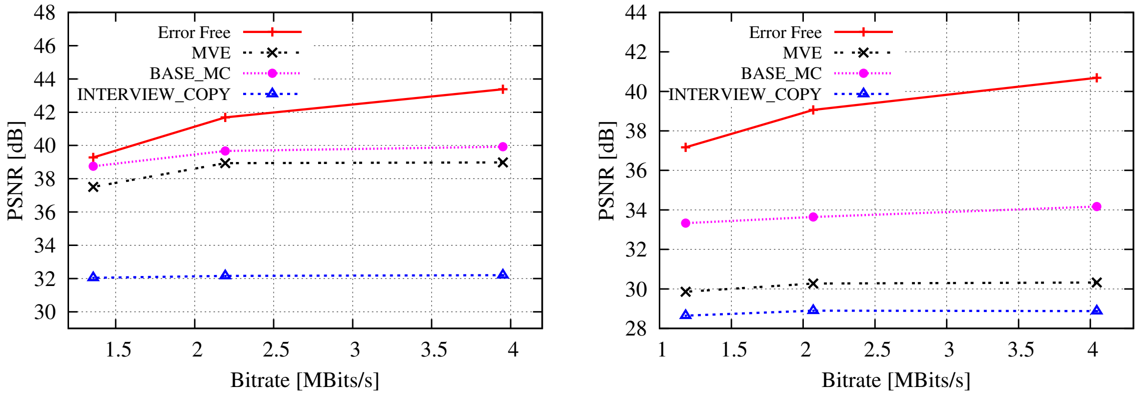
artefacts are present in the reconstructed frame with the BASE_MC method, which result from the average of the three top-left MVs, that were wrongly recovered.

In Figure 6.6 the evaluation of the INTERVIEW_COPY method is also presented. This method achieves lower performance than the other methods for most of the studied cases. This indicates that disparity compensation of pixels of the co-located base view frame is not as efficient as for the BASE_MC method, as it is more dependent on the disparity accuracy. Thus, the noise and errors on the disparity map lead to several artefacts in the reconstructed frame when the INTERVIEW_COPY method is applied. The sequence Champagne Tower, whose results are presented in Figure 6.6(d), is characterised by small motion and high texture complexity, so the difference between successive frames is small and the difference between neighbour pixels within each frame is high. Therefore, small errors in the disparity map have a strong negative effect on the quality of the reconstructed frame, since the neighbour pixels have high differences. Nevertheless, the performance of the motion based methods (MVE and BASE_MC) decreases when high motion is presented. This happens in the sequence Book Arrival and Kendo at frame 46, which results are illustrated in Figure 6.8. It is shown by the results that the INTERVIEW_COPY method maintains a similar performance, achieving higher quality in the reconstructed frame than the other methods.



(a) Book Arrival (b) Kendo

Figure 6.8: Performance evaluation of the different concealment schemes when a P-Frame (frame 46) is lost in the sequences Book Arrival and Kendo.



(a) Balloons (b) Book Arrival

Figure 6.9: Performance evaluation of the different concealment schemes when a B-Frame (frame 45) is lost in the sequences Balloons and Book Arrival.

Figure 6.9 shows the results for two sequences, when a B-frame is lost. The motion and texture complexity for the frame 45 is quite similar to the frame 46 (see Figure 6.4 and 6.5), since they are temporal adjacent frames. However, the performance of the concealment methods is quite different, especially for methods MVE and BASE_MC. This difference occurs due to the accuracy of the source motion vectors used to reconstruct the missing motion field, which are able to better characterise the lost motion information. When using the MVE method, if a B-frame is lost vectors are extrapolated from the temporally close frame in the future, whose MVs point to a frame that precedes the missing one, intercepting the lost frame. This makes the extrapolated vectors more accurate, resulting in reconstructed frame with higher quality. When using the BASE_MC, the disparity compensated vectors come from a B-frame in the base view. Since the B-frames have more prediction modes and references frames (past and future), it is expected that those frames use more motion compensated predictions than the P-frames. Since there are more MVs to use in the concealment, the performance of the BASE_MC increases, as it is able to recover a more accurate motion field for the missing frame.

6.2.3 Conclusions

Summarising, the analysis of the previous results indicates that all methods cope with lost frames and manage to recover the missing data with acceptable quality. Nevertheless, the BASE_MC method revealed to be the more efficient one to recover the lost motion field, as it achieves the higher quality in most of the performed tests. The comparison of the disparity map projection's performance, presented in Table 6.2, revealed that the MVs from the base view are more suitable and lead to higher quality than the extrapolated MVs from the auxiliary view. The results presented in the Table 6.3 indicate that the usage of the average motion vectors to deal with gaps in the reconstructed motion/disparity field does not achieve the desired performance. Therefore, a combination of different techniques can improve the quality of the reconstructed frames. Concluding, the combination of the BASE_MC method, using the base view motion vectors to temporally compensate (motion compensation) the disparity map, with other one (*e.g.*, MVE) is expected to achieve good performance. Consequently quality gains in the reconstruction frame are achieved, comparing with the common methods presented in the literature.

6.3 Proposed concealment method based on a combined motion field

In this section a novel error concealment method to reconstruct lost frames in stereoscopic video decoders based on H.264/MVC is proposed. This method combines motion and disparity information to estimate the motion field for the lost frame, which is used to reconstruct the missing frame itself. An improvement over previous methods in the literature is achieved by reconstructing the disparity map for the missing frame with the motion vectors available in the error-free frames. The motion field is determined using the MVs present in the error-free view, as well as those of previously received frames in the same view. Since the adjacent view mostly provides motion information rather than pixel values, illumination changes and occlusions do not significantly affect the reconstruction of the missing frame, as in previous methods.

Figure 6.10 shows a lost frame, f_t^1, in the auxiliary view (view 1), which is recovered based on the motion information from the base view (mv_t^0), and from the previous frame of view 1 (mv_{t-1}^1). A complete motion field is determined for the missing frame, in order

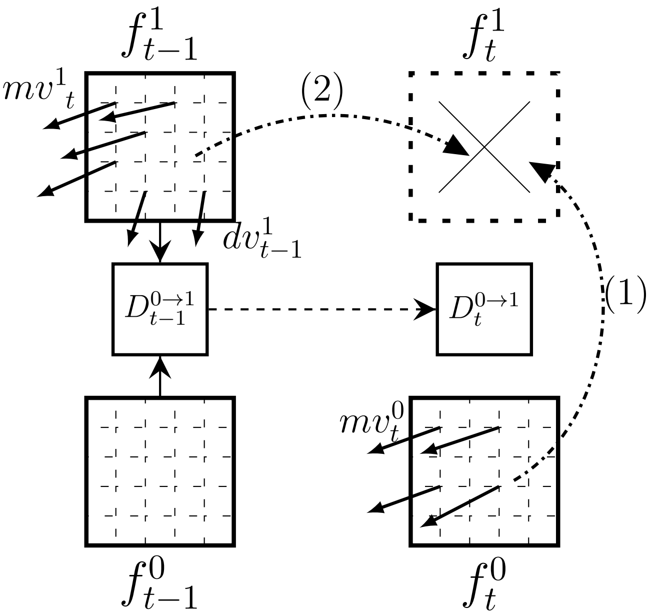


Figure 6.10: Representation of the information used to reconstruct the missing frame using the COMB_MC method.

to reconstruct the missing frame from previously received frames in the same view. This motion field is obtained as follows:

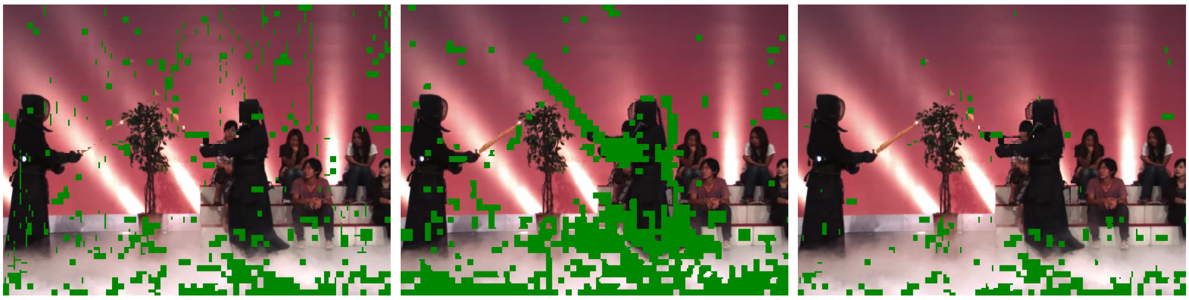
1. The disparity map $D^{0\to1}_{t-1}$ is estimated for the last decoded stereo pair. The disparity map $D^{0\to1}_t$ for the missing frame's instant t is then estimated, by extrapolating $D^{0\to1}_{t-1}$ into time instant t using the mv^0_t.

2. The motion field of the missing frame f^1_t is then determined, by using the MVs associated with the temporally co-located frame in the base view, mv^0_t, compensated through the disparity map $D^{0\to1}_t$, previously estimated. This corresponds to the BASE_MC scheme using the Equation 6.6, and it is illustrated by (1) in Figure 6.10.

3. Each INTRA-coded block in frame f^0_t creates a region with no associated MV. In this case, the MVs obtained from temporal extrapolation of the mv^1_t, are used to fill the gaps. Moreover, the disparity vectors (dv^1_t) present in f^1_{t-1} are also used, as in the MVE concealment scheme, represented by (2) in the figure.

4. Finally, the remaining gaps in motion field of f^1_t, are filled by averaging the MVs of the three (top-left) neighbour blocks.

A novel aspect of this method, referred to as COMB_MC, in comparison with previous ones presented in this chapter, is the use of MVs from both views. Therefore, more motion information is available to reconstruct the missing motion field, avoiding the problem, identified in the Sub-section 6.2.2, related with the unfilled regions that are recovered



(a) CIR= 90.01% - PSNR= 29.59dB (b) CIR= 78.85% - PSNR= 37.56dB (c) CIR= 95.13% - PSNR= 32.54dB

Figure 6.11: Comparison of the proposed method, COMB_MC, (c) with the MVE (a) and BASE_MC (b).

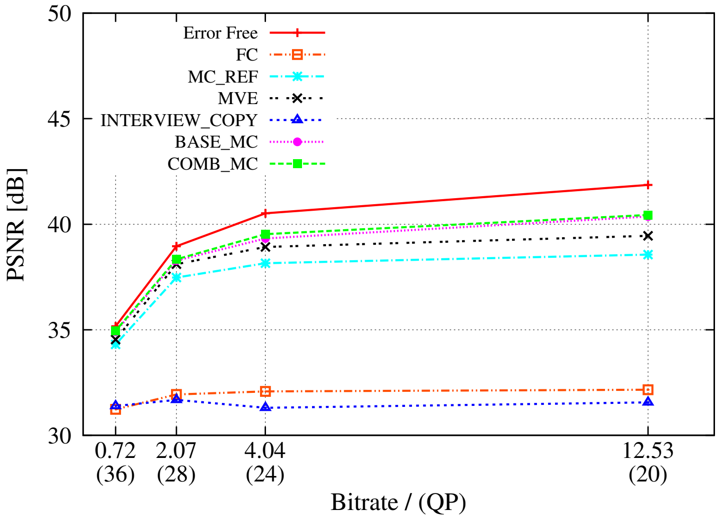
using the neighbours MVs average. Such combination of two motion sources is expected to enhance the method's performance. Moreover, since the reconstruction of the missing frame in the auxiliary view mostly relies on error-free MVs of the base view, it allows for more accurate motion information.

6.3.1 Evaluation of the COMB_MC method

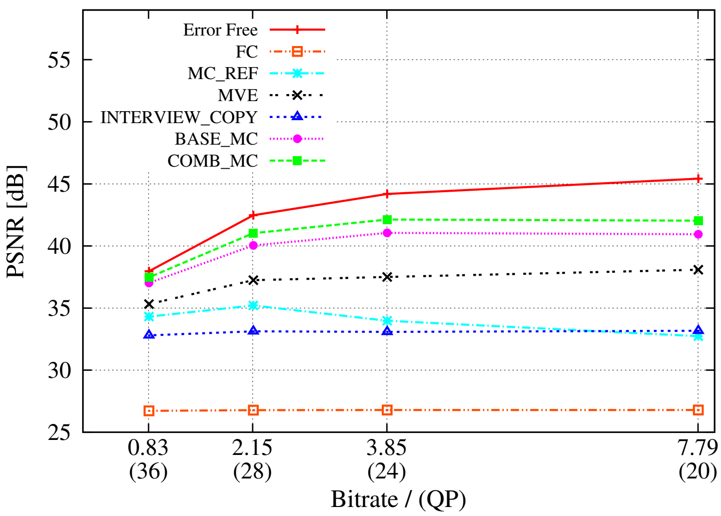
The performance of the joint concealment algorithm is evaluated with existing methods in this section. The tests used the same experimental setup presented in the Sub-section 6.2.1.

The proposed method (COMB_MC), was evaluated against five other methods: frame-copy (FC), that replaces the missing frame with the previously decoded frame in the same view sequence; the motion-copy method (MC_REF), implemented in this work for the auxiliary view in the JM reference software [135], as proposed in [76]; motion recovery only using motion vector extrapolation (MVE), as described in Section 6.1.1 and the BASE_MC described in Sub-section 6.1.2, that only uses disparity compensated motion vectors from the corresponding frame of the base view and the INTERVIEW_COPY method described in the Sub-section 6.1.3.

Firstly the analysis of the motion sources used in the devised method is presented. Figure 6.11 shows the reconstructed frame using only the motion vectors, excluding the top-left neighbour average, for the methods: MVE, BASE_MC and COMB_MC. Thus, the performance of the two motion source, *i.e.*, base and auxiliary view error-free frames, can be evaluated. It is presented, for different schemes, the concealed image ratio (CIR) and the PSNR, calculated only for the recovered regions. The results show that the recovered motion field using the BASE_MC method achieves higher performance than the MVE. Moreover, the COMB_MC is able to recover more regions on the lost motion field (95.13%) against BASE_MC (78.85%) and MVE (90.01%), showing the advantage of



(a) Book Arrival

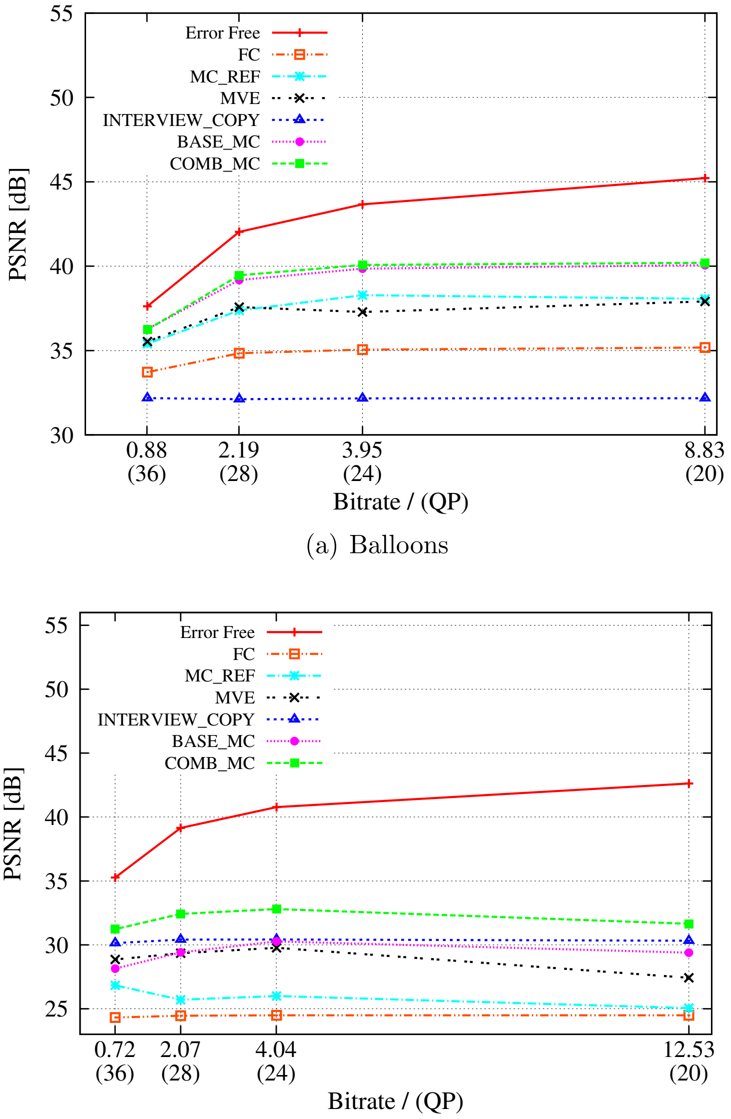


(b) Kendo

Figure 6.12: Quality of the reconstructed B-frame (frame 5) using different concealment methods for different rates ($Mbits/s$).

using two motion sources in the concealment process. Note that, despite of COMB_MC achieving lower performance than the BASE_MC method, it reconstructs more 16% of the image.

Secondly the results of a single frame loss were analysed for two relevant cases, corresponding to quite different cases of motion field recovery: (i) loss of a B frame and (ii) loss of a P frame. Figures 6.12 and 6.13 present the objective quality (PSNR) for the Luma component of the reconstructed frame, achieved by the proposed algorithm and the other five assessed methods. The PSNR of the error-free frame is also presented, as a reference for comparison. These results are shown for different average bitrates ($Mbits/s$) obtained at four different values of QP (20, 24, 28, 36).



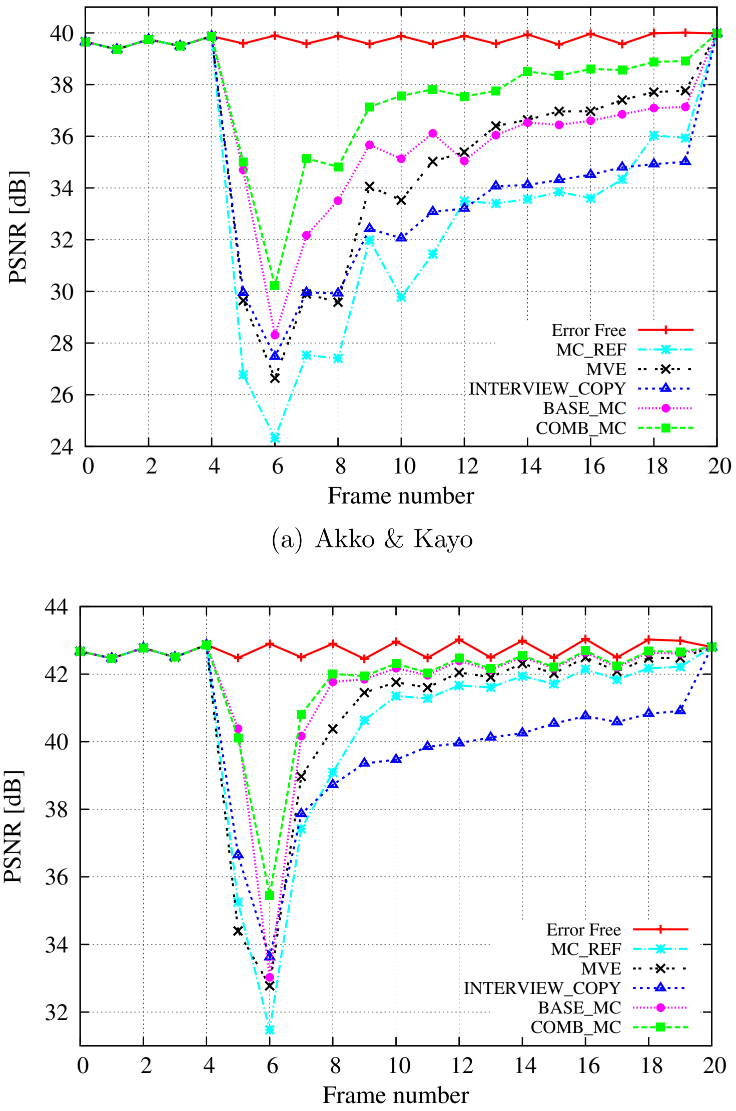
(a) Balloons

(b) Book Arrival

Figure 6.13: Quality of the reconstructed P-frame (frame 18) using different concealment methods for different rates ($Mbits/s$).

Figure 6.12 shows the results when a B frame is lost (frame 5), for two test sequences. In this case, the proposed method provides consistently better results than MC_REF and MVE for all sequences. The proposed concealment method achieves similar results to the BASE_MC scheme. These results show that most of the motion vectors used in error concealment are obtained from the base view frame, demonstrating the usefulness of the disparity information to reconstruct the lost frame, as already shown in this chapter.

Figure 6.13 presents the results of a P frame loss, which is a reference frame. Once again the proposed method clearly outperforms both the MC_REF and the MVE methods. In the case of a P-frame loss, the isolated use of the disparity compensated motion vectors (BASE_MC) is not consistently better than the use of motion vectors from the previous



(a) Akko & Kayo

(b) Book Arrival

Figure 6.14: Error propagation when the frame 6 is lost (P-Frame).

frame. The results for sequence Book Arrival, presented in Figure 6.13(b), show that results of BASE_MC and MVE are worse than results of the INTERVIEW_COPY method. Nevertheless, one may observe that again the COMB_MC method is able to reconstruct the lost frame with a better quality than that achieved by all other methods, for all tested frames.

Since the loss of a reference frame affects the subsequent frames, due to temporal dependencies and prediction mismatch, the results for error propagation were analysed for an entire GOP. Figure 6.14 presents the PSNR obtained along the first GOP of sequence Akko & Kayo and Book Arrival, when frame 6 is lost. This figure shows that methods based on base view motion vectors, *i.e.*, BASE_MC and COMB_MC, outperform all others, achieving a better quality in the concealed frame and leading to higher

Table 6.4: Quality of the reconstructed video sequence, evaluated with the average PSNR (dB) of the auxiliary view versus bitrate (Mbits/s), for different random error patterns with 15% of frame loss and an average burst length of 4 frames.

Sequence	Bitrate	EF	MC_REF	COMB_MC	Gain
Akko & Kayo	3.448	42.16	35.19	36.98	1.79
	1.946	39.81	33.84	35.69	1.86
	1.160	37.08	32.26	33.80	1.54
Balloons	3.951	43.55	39.19	40.81	1.62
	2.194	41.85	38.15	39.57	1.42
	1.358	39.66	36.80	37.96	1.16
Book Arrival	4.045	40.73	35.93	37.52	1.59
	2.071	39.11	35.19	36.39	1.21
	1.182	37.21	34.09	35.06	0.97
Champagne Tower	4.898	43.56	39.00	40.88	1.89
	2.588	41.76	37.88	39.58	1.70
	1.552	39.54	36.37	37.91	1.54
Kendo	3.852	44.45	38.80	40.50	1.71
	2.158	42.80	38.11	39.62	1.51
	1.307	40.72	37.03	38.20	1.17

objective quality in subsequent frames. For the sequence Book Arrival, presented in Figure 6.14(b) the combination of the two motion sources does not improve the reconstructed frame quality. Nevertheless, the results for the sequence Akko & Kayo, presented in Figure 6.14(a), show the effectiveness of the combination of the two concealment methods (MVE and BASE_MC) in the COMB_MC. Such combination increases the motion sources that are available to recover the missing motion field, improving the quality of the reconstructed video sequence from 1 to 4 dB, compared with the isolated used of the MVE and BASE_MC methods.

To validate the effectiveness of the proposed method under more realistic conditions of frame losses in the auxiliary view, further tests were performed using 5 random error patterns (15% frame loss), with an average burst size of 4 frames. Table 6.4 presents the average PSNR obtained with this joint motion method (COMB_MC), compared with the error-free case (EF) and the MC_REF method. The last column presents the PSNR difference between the proposed method and MC_REF. The proposed method consistently outperforms the conventional algorithm for all tested sequences and all tested rates, having a maximum gain of 1.89 dB @ 4.898 $Mbits/s$ (QP=24), for sequence Champagne Tower. The results also show that the advantage of the proposed method increases with the increasing of the bitrate of the compressed sequence. The results demonstrate that the proposed method is an efficient alternative to MC_REF, not only for single events but also for error bursts.

6.4 Joint motion and disparity compensated method

In this section a joint concealment method to reconstruct lost frames in stereoscopic video is presented, to improve the quality of the reconstructed frame achieved with method COMB_MC. The concealment method present in the Section 6.3 combines the two motion information sources available in the error-free frames, *i.e.*, the base and auxiliary view motion vectors. The results shown that the method presented good performance when compared with well-known concealment methods presented in the literature.

An improvement to the COMB_MC method is presented in this section, by introducing more disparity compensated blocks from the base view to reconstruct the missing frame. To achieve this, two concealment schemes presented in Section 6.1 are combined: BASE_MC and INTERVIEW_COPY. The INTERVIEW_COPY concealment scheme is able to achieve similar performance for high and low motion. Therefore, using disparity compensated pixels from the base view may improve the quality of the reconstruct frame in the auxiliary view when high motion is presented.

This method firstly estimates a set of motion vectors $mv_t'^1$ for the missing frame f_t^1 by extrapolating the MVs, mv_{t-1}^1, associated with the last successfully decoded frame in view 1. Those MVs are used to estimate the disparity map associated with the auxiliary view frame ($D_t^{1\to0}$) for the missing frame's instant, through motion compensation. Moreover, another disparity map associated with the base view frame ($D_t^{0\to1}$) is estimated, using the same process as in COMB_MC method. Figure 6.15 illustrates the source of information used to reconstruct the missing frame. As shown in the figure, this method relies on two disparity maps, and has three main stages to recover the missing motion/disparity field:

1. Firstly (see arrow (1) in Figure 6.15) the MVs of the temporally co-located frame in the base view, f_t^0, are disparity compensated using the disparity map $D_t^{0\to1}$. As in BASE_MC method, after the compensation the vectors are grid aligned.

2. The disparity map $D_t^{1\to0}$ is used to recover disparity vectors (INTERVIEW_COPY concealment scheme) for the unrecovered regions of the motion field. This process is represented in the figure by the arrow (2). Those disparity vectors are used to fetch the blocks from the base view frame(f_t^0), decoded without errors.

3. The unfilled regions, due to occlusions in the disparity at the frame's boundary are filled using the extrapolated set of vectors (see arrow (3) in the figure).

Finally, the remaining gaps in motion field of f_t^1, are filled by averaging the MVs of the three (top-left) neighbours. This method extends beyond the COMB_MC concealment method, as it uses the disparity compensated blocks of the base view, before using the

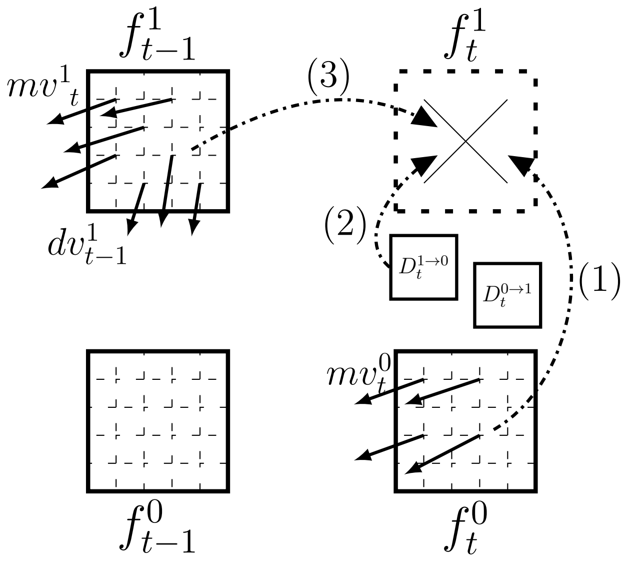


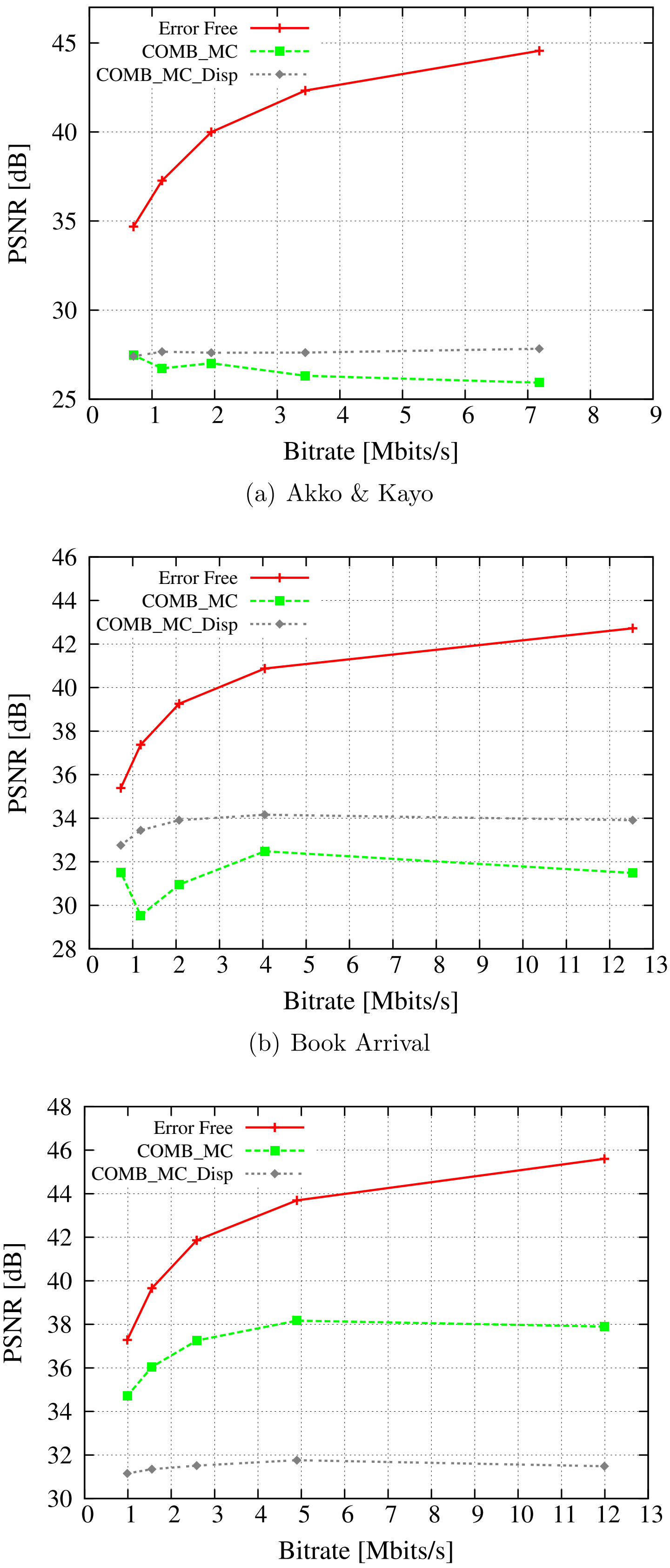
Figure 6.15: Representation of the information used by the COMB_MC_Disp method to reconstruct the missing frame using the .

extrapolated motion vectors from the auxiliary view. In spite of disparity compensated blocks did not show higher performance in the recovery of the whole missing frame, in this method only the unfilled regions, due to INTRA block in the base view, are filled with the base view pixels. This corresponds to those areas of the frame that normally suffer strong temporal changes. Thus, it is expected to achieve higher performance with the INTERVIEW_COPY method, since the pixel information for those regions can be more easily recovered using the base view pixels.

6.4.1 Evaluation of the COMB_MC_Disp method

The performance of the joint motion and disparity concealment algorithm, referred to as COMB_MC_Disp, is evaluated against joint method proposed in the Section 6.3, the COMB_MC. Since that method was already compared with the existing ones, this section only focuses on the comparison of these two methods. The experimental setup is the same presented in the Sub-section 6.2.1.

The quality of the reconstructed frames using the COMB_MC_Disp is evaluated against the COMB_MC method for different single frame loss events and for several stereoscopic sequences. As can be seen in Figure 6.16 the experimental evaluation shows that the COMB_MC_Disp method is able to achieve higher performance than the COMB_MC for sequences with more motion, *i.e.*, Akko & Kayo and Book Arrival. Furthermore, these gains increases with the increasing intensity of the sequence's motion, which is demonstrated by the difference between the results of the sequence Akko & Kayo (higher motion) and Champagne Tower (lower motion), illustrated in Figure 6.16 (a) and (c), re-



(a) Akko & Kayo

(b) Book Arrival

(c) Champagne Tower

Figure 6.16: Comparison between the two proposed concealment methods for different sequences when the frame 46 is lost.

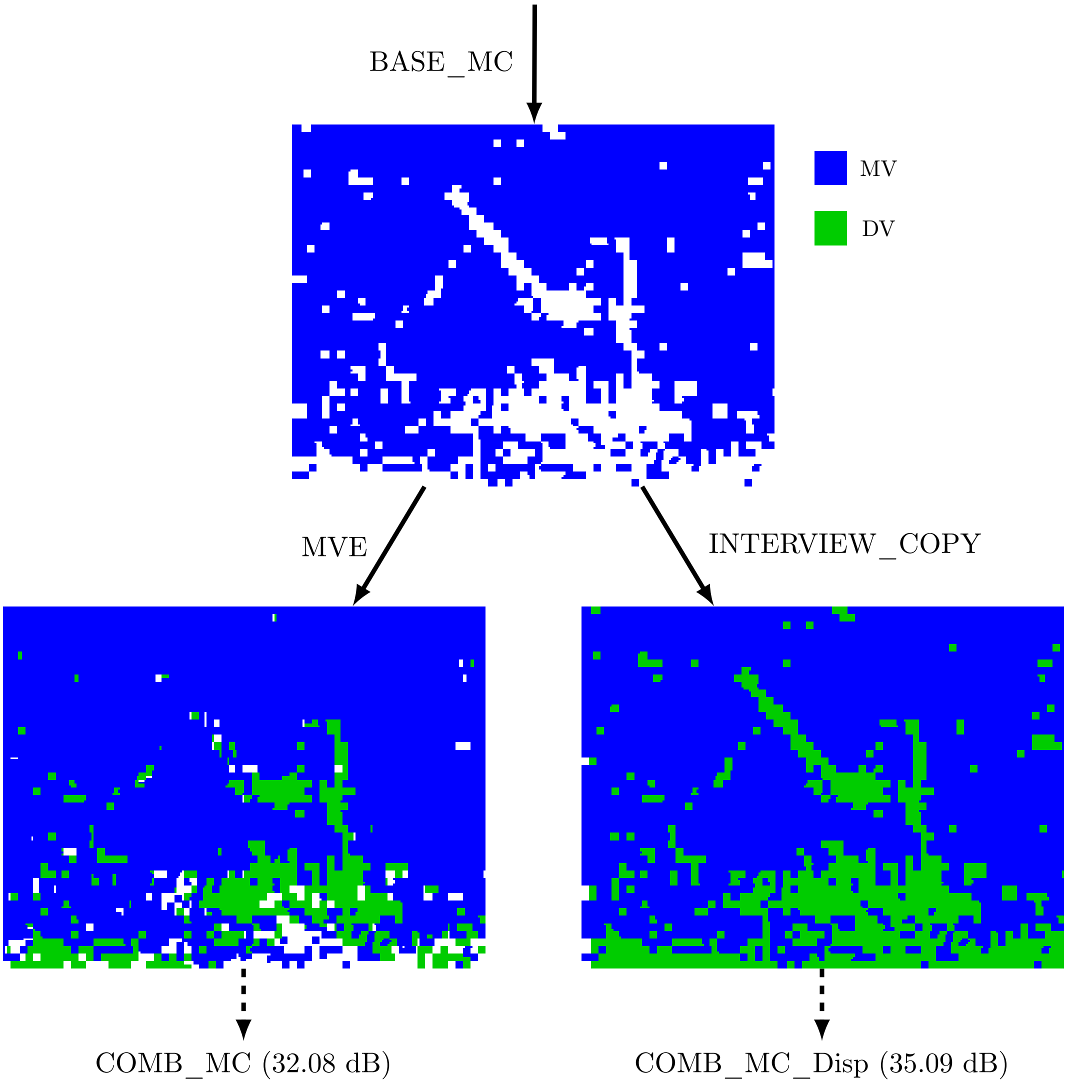
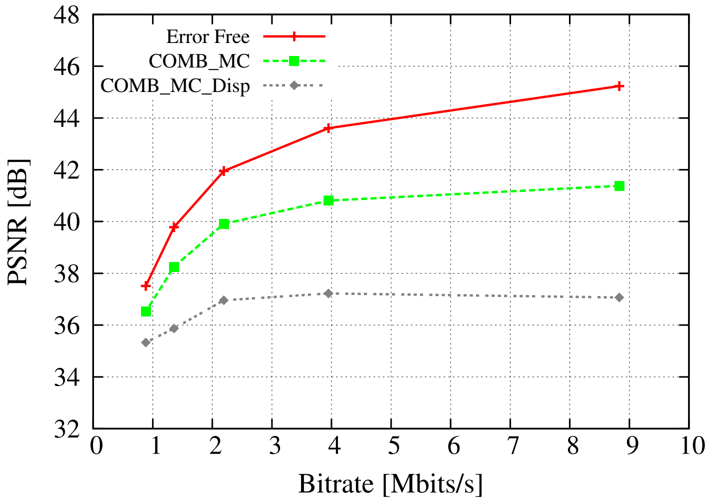


Figure 6.17: Comparison of the reconstructed motion field using the two proposed methods (COMB_MC and COMB_MC_Disp) when the frame 18 of the sequence Kendo is lost.

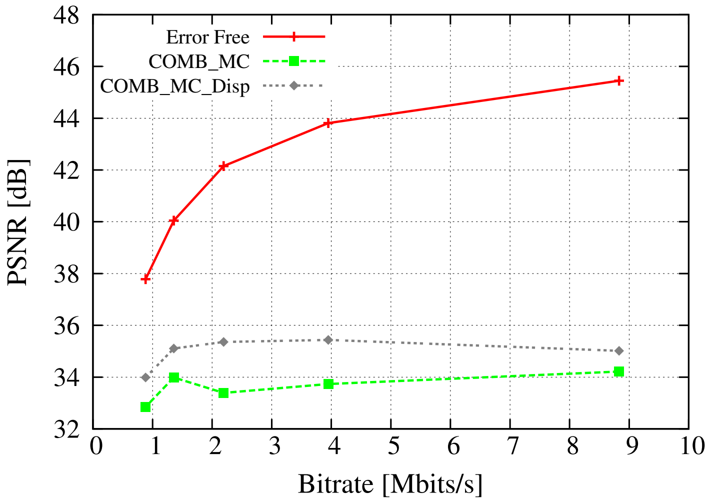
spectively. A different result is achieved for the sequence Champagne Tower, presented in Figure 6.16(c), where the COMB_MC concealment method reconstructs the missing frame with higher objective quality. This occurs due to the higher complexity of the sequence's texture (see Figure 6.5), which decreases the performance of the INTERVIEW_COPY method and consequently the overall performance of the COMB_MC_Disp.

Figure 6.17 presents two initial stages of the two proposed concealment methods described in this chapter. The figure illustrates the type of vectors (*i.e.*, motion or disparity) used to reconstruct the missing motion field, and the white regions are unfilled areas in the missing frame. This figure only presents the centre of the sequence Kendo, and a complete frame is illustrated in Figure A.11. The first stage of the both methods corresponds to the BASE_MC method, *i.e.*, the base view motion vectors are disparity compensated to fill the lost motion field in auxiliary view frame. The unfilled blocks in the missing frame corresponds to gaps in the reconstructed motion field, caused by INTRA blocks in the base view frame. The BASE_MC is able to reconstruct a considerable number of blocks, using motion vectors, but a diagonal region, corresponding to a moving object, is unrecovered. However, the proposed concealment methods use different motion sources, which are able to cope with the unfilled regions, as shown in the bottom images of Figure 6.17. The COMB_MC method uses the motion and disparity vectors presented in the closest frame of the auxiliary view to fill the white blocks. These vectors are not able to accurately recover the information for the unfilled regions, since COMB_MC mostly use MVs that correspond to background regions in the reference frame, so they do not characterise the motion of the moving object. In contrast, the COMB_MC_Disp performs a disparity compensation of the corresponding co-located pixels in the base view frame, using DVs from the disparity map previously obtained (INTERVIEW_COPY method). Since the variations of the disparity map are smaller than the variations of the motion field, the disparity map is able to provide motion information for the unfilled regions that is more accurate than the motion vectors presented in the closest auxiliary view frame. Moreover, the disparity map provides disparity vectors with small variations between neighbouring blocks, therefore smooth regions are obtained with those vectors, providing higher subjective quality.

In Figure 6.18 results for the sequence Balloons are presented for two different frame loss events. Although both of the lost frames correspond to a P-frame (reference), at frame 92 more motion activity is present. Results show that for higher motion the concealment method that uses the disparity compensated pixels from the co-located frame in the base view frame achieves higher quality. This indicates that the disparity vectors can be more accurate than the motion ones present in the auxiliary view in a high motion scenario. Thus, they improve the quality of the reconstructed frame when combined with the available motion vectors of the base view frame, provided by the BASE_MC method. However, the quality of the reconstructed frame with the COMB_MC_Disp concealment method is not consistently higher than the quality achieved by the COMB_MC method. The results for the two frame loss events in the sequence Book Arrival show that the usage of the INTERVIEW_COPY method may not be useful for all test conditions. Consequently, a method should be used to decide when to use the disparity map to fetch the pixels for the auxiliary missing frames from the corresponding pixels of the base view frames.



(a) Lost frame 6



(b) Lost frame 92

Figure 6.18: Quality of the reconstructed frame for the two proposed methods for two lost events in the sequence Balloons.

6.5 Motion based decision method for concealment

6.5.1 Method Description

The previous sections presented and discussed different concealment techniques to cope with whole frame loss in the auxiliary view of stereo video coded with H.264/MVC codec. Two concealment methods were presented and the performance of those methods was evaluated for different frame loss events, showing good results compared with existing concealment methods, as well as simple concealment strategies.

The results presented in previous Sections 6.3 and 6.4 showed that the proposed concealment methods are able to efficiently recover the missing motion/disparity field for

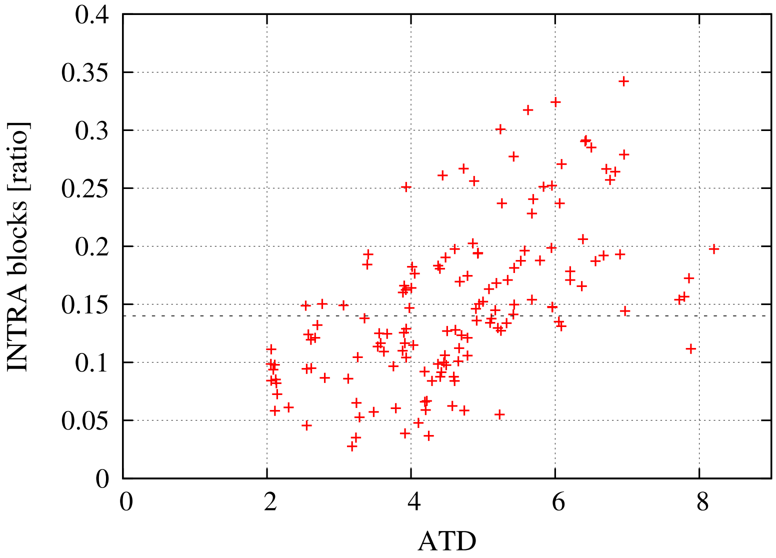


Figure 6.19: Relation between the ratio of the INTRA block in the base view frames and the absolute temporal difference (ATD) for the P coded frames.

the missing frame, which are then used to recover the frame itself. Nevertheless, the performance gain of the concealment methods is not consistent for all frame loss events. For high motion scenarios the COMB_MC_Disp achieves the highest quality in the reconstructed frame, thus, this seems to be an appropriate concealment strategies to cope with errors in stereoscopic sequences with high amplitude moving objects. In contrast, for low motion sequences (*e.g*, Champagne Tower), COMB_MC method achieves higher quality, since it uses less disparity vectors to reconstruct the missing frames when compared with the COMB_MC_Disp method. In order to adapt the concealment strategy to the sequence characteristics, the prediction modes used at the encoder were analysed. In [94] the relation between spatial and temporal predictions in the decoded frames was investigated, in order to choose between different concealment methods, achieving improvements in the quality of the reconstructed frames for different sequences.

The error concealment methods proposed in this chapter aim to cope with lost frames in the auxiliary view of the 3D video, so, the base view frames, which are always available, are expected to provide more reliable information to estimate the error concealment method to reconstruct the missing frames. In Figure 6.19, the relation between the ratio of INTRA blocks in the base view frames and the ATD metric is presented. The ATD value was presented before to measure the temporal differences between frame, and consequently the amount of motion of the stereoscopic video sequences. Thus, the results of figure shows a relation between the number of INTRA blocks and the sequences' motion. The H.264 encoder chooses the prediction modes aiming the best rate-distortion ratio, maintaining a relation between the outputted bits and the reconstructed video quality. The analysis of the chosen predictions (see Figure 6.19) reveals that the encoder increases

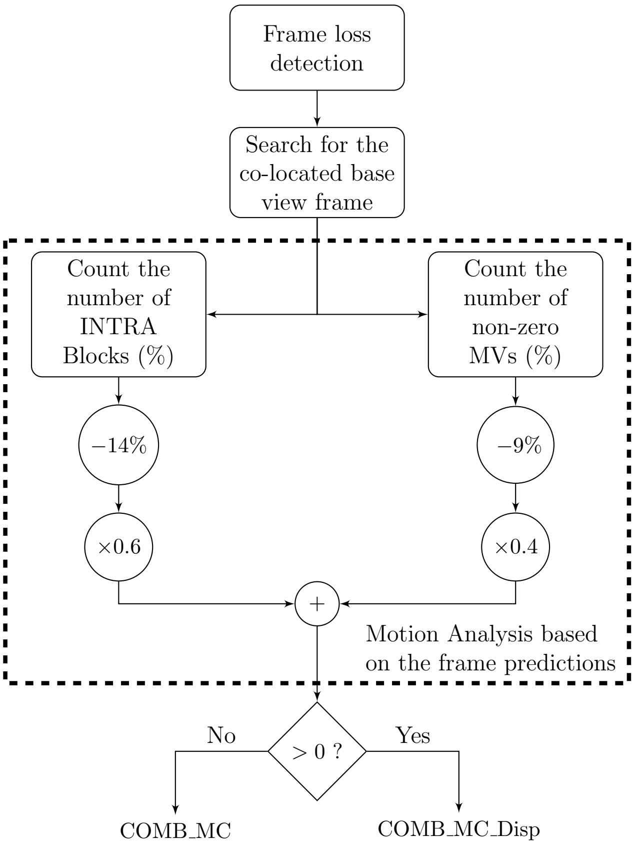

Frame loss detection
Search for the co-located base view frame
Count the number of INTRA Blocks (%)
Count the number of non-zero MVs (%)
−14%
−9%
×0.6
×0.4
+
Motion Analysis based on the frame predictions
No
> 0 ?
Yes
COMB_MC
COMB_MC_Disp

Figure 6.20: Flowchart of the concealment method decision based on the base view frame predictions (dashed square).

the number of the INTRA blocks for the P coded frames, as the amount of motion in the video sequences increases, since it cannot find a suitable prediction using the available reference frames.

Additionally, the number of motion predicted blocks (INTER blocks) also reveals the motion of a video sequence, as the MVs express the motion direction and intensity. Therefore, the number of blocks coded using INTER predictions was also studied in order to achieve a more efficient decision on the concealment method to apply to the missing frames. However, the total amount of MVs used to encode a particular frame may lead to incorrect decisions, when a higher number of zero MVs are presented. For example, when all MVs are zero, although there is a high number of INTER blocks, there is no motion in the video sequence. To overcome this problem, only the number of non-zero MVs was considered to be combined with the number of coded INTRA blocks. This way, when the number of non-zero motion vectors increases, the motion of the video sequence is expected to increase, so the COMB_MC_Disp concealment should be used for this frames.

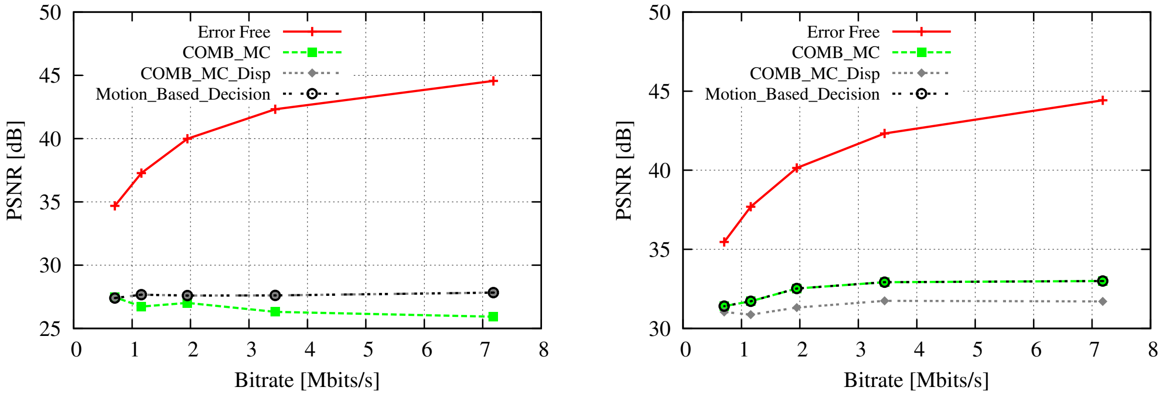
In Figure 6.20, the flowchart of the proposed method to decide which concealment method should be used to recover the missing frames in the auxiliary view is presented. The first operation searches for the co-located base view frame at the missing instant. This frame is used to provide the prediction information used to choose the best concealment method, *i.e.*, the number of the INTRA blocks and the number of non-zeros MVs. As the total number of the blocks varies for different video sequences, a normalised value is used for this purpose. Moreover, the number of non-zero MVs is normalised by the number of INTER blocks, to avoid redundant information and to obtain high values. Thus, in the decision process is always considered percentage values, as shown in the figure.

The results of Figure 6.19 showed that higher motion sequences presented a higher number of INTRA blocks, and a middle threshold can be established at 14% (also shown in the figure). Therefore, the value of number of INTRA blocks is subtracted by 14%, in order to have a number around zero. This indicates that for values higher than zero the COMB_MC_Disp method is applied, otherwise the method COMB_MC is used. The same approach is used to the ratio of non-zero MVs, but in this case the value is subtracted by 9%, value obtained from the analysis of the experimental results. Finally, this two resulted values are combined, giving the more relevance to the number of INTRA blocks, and the final value is used to decide the concealment strategy, as presented in Figure 6.20.

6.5.2 Performance evaluation

This method aims to decide the use of the proper combined concealment method to cope with the missing frames in the stereoscopic video. The results obtained with this approach are presented as the Motion_Based_Decision method. The performance evaluation of the decision method is compared against the results of the combined method, COMB_MC and COMB_MC_Disp. In order to evaluate the proposed method, various stereo sequences have been tested, as shown in Figures 6.21 to 6.24. These results are for single loss events and the quality presented was obtained from the reconstructed frame. Generally, the Motion_Based_Decision method achieves the same quality as the best of the other two concealment methods, indicating that the method presented in this section is appropriate. The results show that the decision approach is able to choose the best of the two proposed methods, for example using the COMB_MC for the sequence Champagne Tower (lower motion), which results are presented in Figure 6.24.

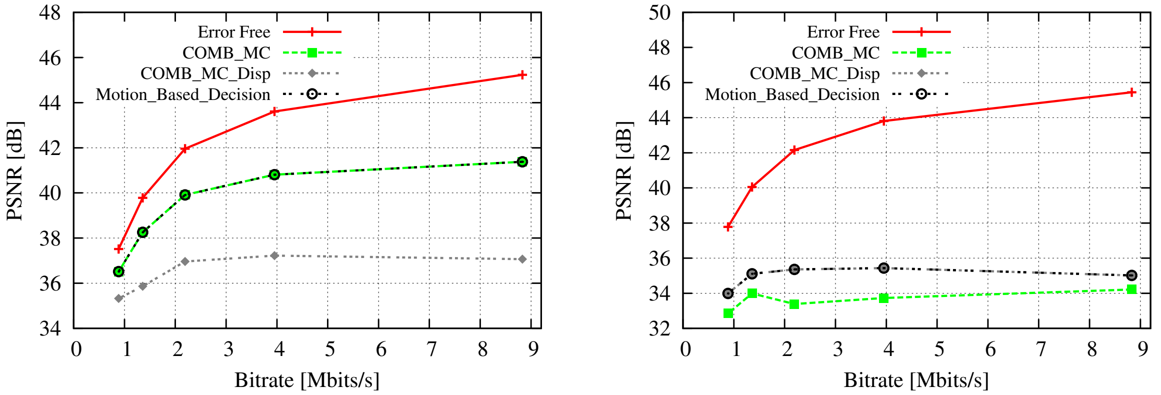
Moreover, in sequences Akko & Kayo, Balloons and Book Arrival the performance of methods COMB_MC and COMB_MC_Disp is not consistent for the different lost events, since the sequences present different amounts of motion in each event. This indicates that the number of INTRA coded blocks provide accurate clues to evaluate the motion of a given sequence. In the sequence Book Arrival (see Figure 6.23) the proposed decision method fail for lower rates. For lower rates, the encoder uses less INTRA blocks, since it can achieve a better ratio-distortion with motion compensated predictions. Consequently, this influences the decision method as the decreasing number of INTRA leads to the usage



(a) Lost frame 46

(b) Lost frame 92

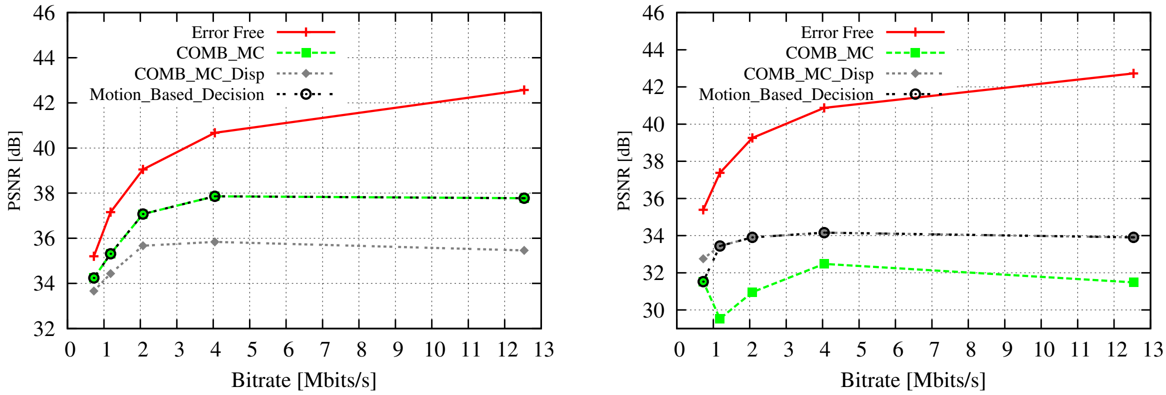
Figure 6.21: Performance evaluation of the decision method for the sequence Akko & Kayo



(a) Lost frame 6

(b) Lost frame 92

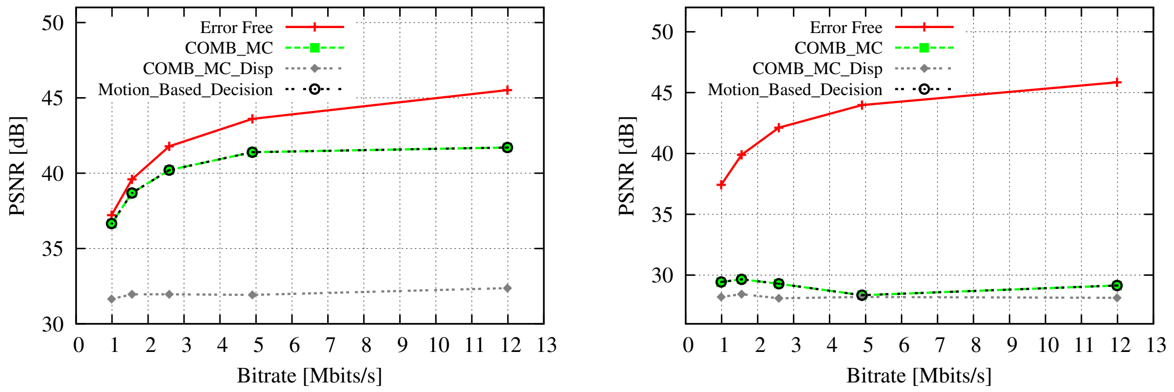
Figure 6.22: Performance evaluation of the decision method for the sequence Balloons.



(a) Lost frame 6

(b) Lost frame 46

Figure 6.23: Performance evaluation of the decision method for the sequence Book Arrival.



(a) Lost frame 6

(b) Lost frame 92

Figure 6.24: Performance evaluation of the decision method for the sequence Champagne Tower.

of the COMB_MC concealment method. Nevertheless, the Motion_Based_Decision is able to take advantage of the highest performances, achieving higher quality in most of the cases. Finally, these results demonstrate, once again, the effectiveness of the information presented in the base view frames to accomplish a efficient concealment in 3D video.

6.6 Summary

The aim of this chapter was to evaluate the performance of different error concealment schemes for 3D video. Those concealment schemes showed to cope with missing frames in the auxiliary view of the 3D video under different conditions, and were able to achieve higher quality than the existing methods. Based on this evaluation study, a combined concealment method (COMB_MC) was proposed and tested under single and random loss events. The results showed that the proposed method is an efficient alternative to motion-copy, avoiding some know issues of the previously proposed methods for 3D video. Subsequently, another combination of the concealment schemes using the disparity compensated pixels from the co-located frame in the base view was developed and evaluated. The results showed that the performance is not consistently for the different types of motion, leading to the development of a motion based decision method. This method revealed good accuracy in selecting the best method for each case. Concluding, the results and discussion presented in this chapter resulted in an efficient frame loss concealment for 3D video based on joint motion/disparity motion field, accomplish the main objective of this chapter.

Chapter 7

Conclusion and future work

This chapter concludes this thesis, presenting some conclusions about this research work and some future research perspectives in the area of 3D video transmission and error concealment.

7.1 Conclusions

This section presents a summary of the of the chapters presented before and a discussion of the contributions presented.

Chapter 2 presented a review on the main technologies for 3D video systems related with the formats for representation, coding methods and proposed storage and transmission systems. This chapter also describes the main features of the MVC extension of the H.264/MVC standard, which is the coding standard for the most 3D applications (*e.g.*, Blu-ray). Moreover, as this research focuses in transmission errors and error concealment, a review of some concealment methods proposed over the last years was presented in the Chapter 3.

The research work presented in this thesis leaded to the development of an advanced 3D video transmission system. More specifically, our research focused on the study of the subjective impacts of transmission errors in 3D video, when the auxiliary view is subject to frame loss. These results were presented in Chapter 4. The results demonstrated that the disparity is intrinsically related with the quality of the concealed 3D video, and that errors have higher effects in those sequences with higher disparity. Moreover, for longer error bursts, the results showed that it is better to switch into 2D visualisation rather than trying to restore the 3D effect by using frame concealment methods.

The results presented in the Chapter 4 leaded to further experiments in order to evaluate the objective quality of the transmitted 3D video using different objective metrics.

A transmission framework was presented in Chapter 5 and the quality evaluation was performed. The quality models, devised from the analysis of the quality results, are a novel contribution to the research community and also for the network planners. Through those models, estimated values of the quality using different metrics can be obtained from the packet loss ratio of the transmission channel. This can be easily measured at the receiver side. The quality models are able to cover different video quality features, as the signal-to-noise relation, the structural similarity and a more specific feature of the 3D video, the stereo perception. The tools developed are part of the first contribution of this research work, which resulted in an efficient 3D receiver to cope with transmission errors in 3D video streams, delivered using the MPEG-2 TS encapsulation. Moreover, as part of the receiver side, an efficient 3D video decoder was achieved, that is able to detect and conceal the missing frames using different concealment techniques. This set of tools provides relevant knowledge, for further investigations in quality evaluation and error concealment.

The second contribution of this research work was the development of efficient concealment methods, able to recover the motion field for missing frames in 3D video. Assuming the hierarchical transmission system, as presented in Chapter 5, the proposed methods are an efficient alternative to the existing ones. They are mainly based on the information presented in the error-free view (*i.e.*, the base view), since these information can provide more accurate clues to estimate the loss data in the auxiliary view, as described in Chapter 6. The presented results showed the effectiveness of the proposed techniques, when compared with conventional motion-based algorithms, achieving an average gain of 1.89 dB in random frame loss and up to 6 dB of gain in the objective quality of the reconstructed frame. Moreover, an improvement to the proposed method was devised, and a motion based decision approach was incorporated, combining both methods, thus allowing to improve the quality of the reconstructed frames up 2 dB.

One may conclude that the research on alternative frame loss concealment techniques may bring new insights to achieve a more efficiently video decoding in the presence of error prone channels. Moreover, the subjective study presented in this dissertation provided relevant subjective factors that may be used in future developments of error concealment techniques. New representation formats for 3D video and novel coding approaches are being proposed, leading to new error concealment problems, and different ways to recovery the missing information, thus, the research work in this area should continue.

7.2 Future work

In this research work, the technologies and the subjective quality impact of errors in 3D video were studied. Nevertheless, there are still a number of topics that require further investigation, and may lead to more efficient 3D video decoders, namely:

- **3D video quality models extension to the channel level.** The 3D video quality models presented in Chapter 5 revealed to match with the experimental data. Thus, the combination with the hierarchical DVB-T channel models could extend the proposed models to a new level, where the quality of the 3D video could be estimated from the channel conditions, instead of the packet loss. This would extend the relevance of the proposed models to the network planners and developers.

- **Extension of the proposed method to a multi-view scenario.** The proposed concealment methods, described in Chapter 6, revealed good performance to conceal missing frames in 3D video (*i.e.*, stereoscopic video). Therefore, the same techniques are expected to have similar behaviour under a multi-view scenario, where, besides the base view frames, other auxiliary view frames that may provide motion information clues to estimate the lost frames' motion field.

- **Extend the concealment algorithm to MVD coding scenario.** Although the proposed concealment methods did not focus in the recovery of the depth information, they could be used to recover texture in multi-view plus depth video signals. In this case, besides disparity information, depth maps are available to extract relevant information, for example, to detect occlusions and errors in the disparity map. The performance of the proposed method could also be investigated to reconstruct the lost depth map.

- **Performance evaluation of the concealment methods in a MDC.** The multi-description coding approach is being investigated over the last years, as a novel alternative for the emergent multi-path networks. The algorithms presented in the proposed method have great flexibility and could be applied under different coding scenarios. Moreover, since MDC introduces new possibilities to error concealment by using the redundant information in each description, such characteristic will be used to investigate how different types of redundancy can be jointly used with the proposed concealment techniques to enhance the quality of received views.

Bibliography

[1] P. Merkle, K. Müller, and T. Wiegand, "3D video: acquisition, coding, and display," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 946–950, May 2010.

[2] D. Broberg, "Infrastructures for home delivery, interfacing, captioning, and viewing of 3-D content," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 684–693, Apr. 2011.

[3] A. Gotchev, G. Akar, T. Capin, D. Strohmeier, and A. Boev, "Three-dimensional media for mobile devices," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 708–741, Apr. 2011.

[4] I. Feldmann, W. Waizenegger, N. Atzpadin, and O. Schreer, "Real-time depth estimation for immersive 3D videoconferencing," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Jun. 2010, pp. 1–4.

[5] A. Vetro, T. Wiegand, and G. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[6] "ISO/IEC JTC 1/SC 29/WG 11, ISO/IEC 23002-3:2007 - MPEG video technologies part 3: Representation of auxiliary video and supplemental information," Doc. N8768, Marrakech, Morocco, Jan. 2007.

[7] G. Akar, A. Tekalp, C. Fehn, and M. Civanlar, "Transport methods in 3DTV - a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1622–1630, Nov. 2007.

[8] S. Wenger, M. M. Hannuksela, T. Stockhammer, M. Westerlund, and S. D., "RFC 3984: RTP payload format for H.264 video," Feb. 2005.

[9] J. Morgade, A. Usandizaga, P. Angueira, D. de la Vega, A. Arrinda, M. Velez, and J. Ordiales, "3DTV roll-out scenarios: A DVB-T2 approach," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 582–592, Jun. 2011.

[10] C. G. Gürler, B. Görkemli, G. Saygili, and A. Tekalp, "Flexible transport of 3-D video over networks," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 694–707, Apr. 2011.

[11] K. Müller, P. Merkle, G. Tech, and T. Wiegand, "3D video formats and coding methods," in *IEEE International Conference on Image Processing*, Sep. 2010, pp. 2389 –2392.

[12] T. Stockhammer, M. Hannuksela, and T. Wiegand, "H.264/AVC in wireless environments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 657–673, Jul. 2003.

[13] B. Micallef and C. Debono, "An analysis on the effect of transmission errors in real-time H.264-MVC bit-streams," in *IEEE Mediterranean Electrotechnical Conference*, Apr. 2010, pp. 1215–1220.

[14] K. Wang, M. Barkowsky, K. Brunnstrom, M. Sjostrom, R. Cousseau, and P. Le Callet, "Perceived 3D TV transmission quality assessment: Multi-laboratory results using absolute category rating on quality of experience scale," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 544–557, Dec. 2012.

[15] N. Hur, H. Lee, G. S. Lee, S. J. Lee, A. Gotchev, and S.-I. Park, "3DTV broadcasting and distribution systems," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 395–407, Jun. 2011.

[16] G. Tech, A. Smolic, H. Brust, P. Merkle, K. Dix, Y. Wang, K. Müller, and T. Wiegand, "Optimization and comparision of coding algorithms for mobile 3DTV," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2009, pp. 1–4.

[17] L. Jiang, J. He, N. Zhang, and T. Huang, "An overview of 3D video representation and coding," *3D Research*, vol. 1, pp. 43–47, 2010.

[18] A. Vetro, A. Tourapis, K. Müller, and T. Chen, "3D-TV content storage and transmission," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 384–394, Jun. 2011.

[19] A. Vetro, "Representation and coding formats for stereo and multiview video," in *Intelligent Multimedia Communication: Techniques and Applications*, ser. Studies in Computational Intelligence, C. Chen, Z. Li, and S. Lian, Eds. Springer Berlin/Heidelberg, 2010, vol. 280, pp. 51–73.

[20] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, “An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution,” in *Picture Coding Symposium (PCS)*, May 2009, pp. 1–4.

[21] P. Merkle, H. Brust, K. Dix, K. Müller, and T. Wiegand, “Stereo video compression for mobile 3D services,” in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2009, pp. 1–4.

[22] G. Sullivan and T. Wiegand, “Video compression - from concepts to the H.264/AVC standard,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 18–31, Jan. 2005.

[23] H. Kalva, “The H.264 video coding standard,” *IEEE Multimedia*, vol. 13, no. 4, pp. 86–90, Dec. 2006.

[24] ITU and ISO/IEC JTC1, “ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Advanced video coding for generic audiovisual services,” 2010.

[25] A. Vetro, “Frame compatible formats for 3D video distribution,” in *IEEE International Conference on Image Processing*, Sep. 2010, pp. 2405 –2408.

[26] C. Latry and B. Rouge, “Super resolution: quincunx sampling and fusion processing,” in *IEEE International Geoscience and Remote Sensing Symposium*, vol. 1, Jul. 2003, pp. 315–317.

[27] G. Ballocca, P. D’Amato, M. Grangetto, and M. Lucenteforte, “Tile format: A novel frame compatible approach for 3D video broadcasting,” in *IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2011, pp. 1–4.

[28] G. J. Sullivan, A. M. Tourapis, T. Yamakage, and C. S. Lim, “Draft AVC amendment text to specify constrained baseline profile, stereo high profile, and frame packing SEI message,” in *Joint Video Team (JVT) Doc. JVT-AE204*, London, U.K, Jul. 2009.

[29] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y.-K. Wang, “Joint draft 8 of multiview video coding,” in *Joint Video Team (JVT) Doc. JVT-AB204*, Hannover, Germany, Jul. 2008.

[30] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, “Efficient prediction structures for multiview video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, Nov. 2007.

[31] K. Müller, P. Merkle, and T. Wiegand, “3-D video representation using depth maps,” *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[32] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. A. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimized approach on 3D-TV," in *Int. Broadcast Conf. 2002*, 2002, pp. 357–365.

[33] S. Chan, H.-Y. Shum, and K.-T. Ng, "Image-based rendering and synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 22–33, Nov. 2007.

[34] P. Ndjiki-Nya, M. Koppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-D video," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 453–465, Jun. 2011.

[35] H. Oh and Y.-S. Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," in *Advances in Image and Video Technology*, ser. Lecture Notes in Computer Science, L.-W. Chang and W.-N. Lie, Eds. Springer Berlin / Heidelberg, 2006, vol. 4319, pp. 898–907.

[36] Y. Morvan, D. Farin, and P. de With, "Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images," in *IEEE International Conference on Image Processing*, vol. 5, Oct. 2007, pp. 105–108.

[37] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. With, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 73–88, 2009.

[38] N. Zhang, S. Ma, and W. Gao, "Shape-based depth map coding," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Sep. 2009, pp. 316–319.

[39] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110, Jan. 2005.

[40] K. Müller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *IEEE Workshop on Multimedia Signal Processing*, Oct. 2008, pp. 34–39.

[41] A. Bourge, J. Gobert, and F. Bruls, "MPEG-C part 3: Enabling the introduction of video plus depth contents," in *IEEE Workshop on Content Generation and Coding for 3D-Television*, 2006, pp. 3–6.

[42] W. Bruls, C. Varekamp, R. Gunnewiek, B. Barenbrug, and A. Bourge, "Enabling introduction of stereoscopic (3D) video: Formats and compression standards," in *IEEE International Conference on Image Processing*, vol. 1, Oct. 2007, pp. 89–92.

[43] "ITU-T and ISO/IEC JTC 1, Final draft, Amendment 3 to ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video)," Doc. N1366, Sep. 1996.

[44] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," *EURASIP Journal on Advances in Signal Processing*, pp. 1–13, 2009.

[45] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[46] N. Jin, F. Li, and X. Lai, "Disparity estimation with disparity field correlation and epipolar geometry constraint for multiview video coding," in *International Conference on Intelligent Control and Information Processing*, vol. 1, Jul. 2011, pp. 260–263.

[47] H.-Y. Yang, J.-Y. Lin, H.-W. Tsao, and Y.-C. Fan, "Algorithm and architecture design of illumination changes adaptive motion estimation," in *IEEE International Symposium on Consumer Electronics*, May 2009, pp. 565–568.

[48] M. Karczewicz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 637–644, Jul. 2003.

[49] Y.-K. Wang, Y. Chen, and M. M. Hannuksela, "Time-first coding for multi-view video coding," in *Joint Video Team (JVT) Doc. JVT-U104,*, Hangzhou, China, Oct. 2009.

[50] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *IEEE International Conference on Multimedia and Expo*, Jul. 2006, pp. 1929–1932.

[51] Y. Chen, Y.-K. Wang, and M. Gabbouj, "Buffer requirement analyses for multiview video coding," in *Picture Coding Symposium (PCS)*, Lisbon, Portugal, Nov. 2007.

[52] T. Schierl and S. Narasimhan, "Transport and storage systems for 3-D video using MPEG-2 systems, RTP, and ISO file format," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 671–683, Apr. 2011.

[53] "ITU-T Rec. H.222.0 and ISO/IEC 13818-1:2007 information technology - generic coding of moving pictures and associated audio information: Systems," May 2006.

[54] B. Lechner, R. Chernock, M. Eyer, A. Goldberg, and M. Goldman, "The ATSC transport layer, including program and system information protocol (PSIP)," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 77–101, Jan. 2006.

[55] T. Schierl, K. Gruneberg, S. Narasimhan, and A. Vetro, "ISO/IEC 13818-1:2007/AMD4 - Transport of multiview video over ITU-T Rec H.222.0 — ISO/IEC 13818-1, ISO/IEC JTC1/SC29/WG11," London, UK, Sep. 2009.

[56] D. Wu, L. Sun, and S. Yang, "A selective transport framework for delivery MVC video over MPEG-2 TS," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Jun. 2011, pp. 1–6.

[57] T. Chen and Y. Kashiwagi, "Subjective picture quality evaluation of MVC stereo high profile for full-resolution stereoscopic high-definition 3D video applications," in *Conference of Signal and Image Processing*, Maui, Hawaii, USA, Aug. 2010.

[58] Y. Wang, S. Wenger, J. Wen, and A. Katsaggelos, "Error resilient video coding techniques," *IEEE Signal Processing Magazine*, vol. 17, no. 4, pp. 61–82, Jul. 2000.

[59] B. Yan, H. Gharavi, and B. Hu, "Pixel interlacing based video transmission for low-complexity intra-frame error concealment," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 253–257, Jun. 2011.

[60] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: a review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[61] P. Salama, N. Shroff, E. Coyle, and E. Delp, "Error concealment techniques for encoded video streams," in *International Conference on Image Processing*, vol. 1, Oct. 1995, pp. 9–12.

[62] S. Valente, C. Dufour, F. Groliere, and D. Snook, "An efficient error concealment implementation for MPEG-4 video streams," *IEEE Transactions on Consumer Electronics*, vol. 47, no. 3, pp. 568–578, Aug. 2001.

[63] J. W. Woods, *Multidimensional Signal, Image, and Video Processing and Coding*, 2nd ed. Orlando, FL, USA: Academic Press, Inc., 2012.

[64] R. Bernardini, L. Celetto, G. Gennari, M. Petrani, and R. Rinaldo, "Error concealment of INTRA coded video frames," in *International Workshop on Image Analysis for Multimedia Interactive Services*, Apr. 2010, pp. 1–4.

[65] W. Kim, J. Koo, and J. Jeong, "Fine directional interpolation for spatial error concealment," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 1050–1056, Aug. 2006.

[66] S.-C. Hsia, "An edge-oriented spatial interpolation for consecutive block error concealment," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 577–580, Jun. 2004.

[67] J. Chen, J. Liu, X. Wang, and G. Chen, "Modified edge-oriented spatial interpolation for consecutive blocks error concealment," in *IEEE International Conference on Image Processing,*, vol. 3, Sep. 2005, pp. 904–913.

[68] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 1153–1156.

[69] H. Senel, "Gradient estimation using wide support operators," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 867–878, Apr. 2009.

[70] M. Kim, H. Lee, and S. Sull, "Spatial error concealment for H.264 using sequential directional interpolation," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1811–1818, Nov. 2008.

[71] W. Zeng and B. Liu, "Geometric-structure-based error concealment with novel applications in block-based low-bit-rate coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 648–665, Jun. 1999.

[72] Y. Wang, Q.-F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *IEEE Transactions on Communications*, vol. 41, no. 10, pp. 1544 –1551, oct 1993.

[73] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Transactions on Image Processing*, vol. 4, no. 4, pp. 470–477, Apr. 1995.

[74] M. Kim, S.-W. Lee, and S.-D. Kim, "Spatial error concealment method based on POCS with a correlation-based initial block," *IET Image Processing*, vol. 1, no. 2, pp. 134–140, Jun. 2007.

[75] S. Hemami and T.-Y. Meng, "Transform coded image reconstruction exploiting interblock correlation," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 1023 –1027, Jul. 1995.

[76] S. K. Bandyopadhyay, Z. Wu, P. Pandit, and J. M. Boyce, "An error concealment scheme for entire frame losses for H.264/AVC," in *Sarnoff Symposium, IEEE*, Mar. 2006, pp. 1–4.

[77] H. Liu, D. Wang, W. Li, and O. Issa, “New method for concealing entirely lost frames in H.264 video transmission over wireless networks,” in *IEEE International Symposium on Consumer Electronics (ISCE)*, Jun. 2011, pp. 112–116.

[78] K. Song, T. Chung, Y. Kim, Y. Oh, and C.-S. Kim, “Error concealment of H.264/AVC video frames for mobile video broadcasting,” *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 704–711, May 2007.

[79] J.-T. Chien, G.-L. Li, and M.-J. Chen, “Effective error concealment algorithm of whole frame loss for H.264 video coding standard by recursive motion vector refinement,” *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1689–1695, Aug. 2010.

[80] Y. Chen, Y. Hu, O. Au, H. Li, and C. W. Chen, “Video error concealment using spatio-temporal boundary matching and partial differential equation,” *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 2–15, Jan. 2008.

[81] S. Garg and S. Merchant, “Interpolated candidate motion vectors for boundary matching error concealment technique in video,” *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 53, no. 10, pp. 1039–1043, Oct. 2006.

[82] T. Thaipanich, P.-H. Wu, and C.-C. Kuo, “Low-complexity video error concealment for mobile applications using OBMA,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 753–761, May 2008.

[83] J. Zhang, J. Arnold, and M. Frater, “A cell-loss concealment technique for MPEG-2 coded video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 4, pp. 659 –665, Jun. 2000.

[84] Y. Sun, S. Sun, X. Jing, and L. Zhao, “A dynamic temporal error concealment algorithm for H.264,” in *2010 International Conference on Multimedia Technology (ICMT)*, Oct. 2010, pp. 1–4.

[85] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, “An error concealment algorithm for streaming video,” in *International Conference on Image Processing*, vol. 3, Sep. 2003, pp. 49–52.

[86] P. Salama, N. Shroff, and E. Delp, “Error concealment in MPEG video streams over ATM networks,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1129–1144, Jun. 2000.

[87] Q. Peng, T. Yang, and C. Zhu, “Block-based temporal error concealment for video packet using motion vector extrapolation,” in *IEEE International Conference on*

Communications, Circuits and Systems and West Sino Expositions, vol. 1, Jul. 2002, pp. 10–14.

[88] H. Liu, W. Li, and O. Issa, "New algorithm for motion vector extrapolation for concealing entire frame loss in the H.264 video receiver," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2011, pp. 525–526.

[89] Y. Chen, K. Yu, J. Li, and S. Li, "An error concealment algorithm for entire frame loss in video transmission," in *Picture Coding Symposium (PCS)*, 2004.

[90] B. Yan and H. Gharavi, "A hybrid frame concealment algorithm for H.264/AVC," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 98–107, Jan. 2010.

[91] J.-W. Suh and Y.-S. Ho, "Error concealment techniques for digital TV," *IEEE Transactions on Broadcasting*, vol. 48, no. 4, pp. 299–306, Dec. 2002.

[92] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, Aug. 1981.

[93] B. Micallef and C. Debono, "Error concealment techniques for multi-view video," in *IFIP Wireless Days (WD)*, Oct. 2010, pp. 1–5.

[94] L. Pang, M. Yu, W. Yi, G. Jiang, W. Liu, and Z. Jiang, "Relativity analysis-based error concealment algorithm for entire frame loss of stereo video," in *International Conference on Signal Processing*, vol. 2, Nov. 2006.

[95] Ç. Bilen, A. Aksay, and G. B. Akar, "Two novel methods for full frame loss concealment in stereo video," in *Picture Coding Symposium (PCS)*, Lisbon, Portugal, 2007.

[96] Y. Chen, C. Cai, and K.-K. Ma, "Stereoscopic video error concealment for missing frame recovery using disparity-based frame difference projection," in *IEEE International Conference on Image Processing*, Nov. 2009, pp. 4289–4292.

[97] T.-Y. Chung, S. Sull, and C.-S. Kim, "Frame loss concealment for stereoscopic video based on inter-view similarity of motion and intensity difference," in *IEEE International Conference on Image Processing,*, Sep. 2010, pp. 441–444.

[98] ——, "Frame loss concealment for stereoscopic video plus depth sequences," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1336–1344, Aug. 2011.

[99] S. Yang, Y. Zhao, S. Wang, and H. Chen, "Error concealment for stereoscopic video using illumination compensation," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1907–1914, Nov. 2011.

[100] W. Miled, J.-C. Pesquet, and M. Parent, “A convex optimization approach for depth estimation under illumination variation,” *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 813–830, Apr. 2009.

[101] I.-L. Jung, T. Chung, K. Song, and C.-S. Kim, “Efficient stereo video coding based on frame skipping for real-time mobile applications,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 3, pp. 1259 –1266, August 2008.

[102] S. Yasakethu, W. Fernando, B. Kamolrat, and A. Kondoz, “Analyzing perceptual attributes of 3D video,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 864–872, May 2009.

[103] “*ITU-T Rec. J.144*, objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” 2004.

[104] S. Winkler, *Digital Video Quality: Vision Models and Metrics.* Wiley, Mar. 2005.

[105] H. Wu and E. K.R. Rao, *Digital Video Image Quality and Perceptual Coding.* CRC Press, 2006.

[106] Q. Huynh-Thu and M. Ghanbari, “Temporal aspect of perceived quality in mobile video broadcasting,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 641–651, Sep. 2008.

[107] ——, “No-reference temporal quality metric for video impaired by frame freezing artefacts,” in *IEEE International Conference on Image Processing*, Nov. 2009, pp. 2221–2224.

[108] I. Dinstein, M. G. Kim, J. Tselgov, and A. Henik, “Compression of stereo images and the evaluation of its effects on 3-D perception,” in *Applications of Digital Image Processing XII, SPIE*, vol. 1153, 1989, pp. 522–530.

[109] S. Pastoor, “3D-television: A survey of recent research results on subjective requirements,” in *Signal Processing: Image Commun.*, vol. 4, no. 1, 1991, pp. 21–32.

[110] M. Perkins, “Data compression of stereopairs,” *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684 –696, Apr. 1992.

[111] A. Benoit, P. L. Callet, P. Campisi, and R. Cousseau, “Quality assessment of stereoscopic images,” in *EURASIP Journal on Image and Video Processing*, 2008.

[112] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo, “Objective quality assessment method of stereo images,” in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2009, pp. 1–4.

[113] L. Xing, J. You, T. Ebrahimi, and A. Perkis, “Estimating quality of experience on stereoscopic images,” in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Dec. 2010, pp. 1–4.

[114] P. Gorley and N. Holliman, “Stereoscopic image quality metrics and compression,” 2008.

[115] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, “Stereo image quality: effects of mixed spatio-temporal resolution,” in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, 2000, pp. 188–193.

[116] C. Hewage, S. Worrall, S. Dogan, S. Villette, and A. Kondoz, “Quality evaluation of color plus depth map-based stereoscopic video,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304–318, Apr. 2009.

[117] G. Leon, H. Kalva, and B. Furht, “3D video quality evaluation with depth quality variations,” in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2008.

[118] Z. Li, J. Chakareski, X. Niu, Y. Zhang, and W. Gu, “Modeling of distortion caused by markov-model burst packet losses in video transmission,” in *IEEE International Workshop on Multimedia Signal Processing*, Oct. 2009, pp. 1–6.

[119] O. Hohlfeld and G. Geib, R.; Hasslinger, “Packet loss in real-time services: Markovian models generating QoE impairments,” in *International Workshop on Quality of Service*, Jun. 2008, pp. 239–248.

[120] “ITU-R Recommendation BT.500-12, Methodology for the subjective assessment of the quality of television pictures,” 2009.

[121] “ITU-R Recommendation P.910, Subjective video quality assessment methods for multimedia applications,” 2008.

[122] “ITU-R Recommendation BT.1438, Subjective assessment of stereoscopic television pictures,” 2000.

[123] J. Carreira, L. Pinto, N. Rodrigues, S. Faria, and P. Assuncao, “Subjective assessment of frame loss concealment methods in 3D video,” in *Picture Coding Symposium (PCS)*, Dec. 2010, pp. 182–185.

[124] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, “Using disparity for quality assessment of stereoscopic images,” in *IEEE International Conference on Image Processing*, Oct. 2008, pp. 389–392.

[125] K. Alajel, W. Xiang, and Y. Wang, "Unequal error protection scheme based hierarchical 16-QAM for 3-D video transmission," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 731–738, Aug. 2012.

[126] C. Hellge, S. Mirta, T. Schierl, and T. Wiegand, "Mobile TV with SVC and hierarchical modulation for DVB-H broadcast services," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–5, May 2009.

[127] M. Barkowsky, K. Wang, R. Cousseau, K. Brunnstrom, R. Olsson, and P. Le Callet, "Subjective quality assessment of error concealment strategies for 3DTV in the presence of asymmetric transmission errors," *International Packet Video Workshop*, pp. 193–200, Dec. 2010.

[128] G. Saygili, C. Gurler, and A. Tekalp, "Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 593–601, Jun. 2011.

[129] Y. Shen, C. Lü, , P. Xu, and L. Xu, "Objective quality assessment of noised stereoscopic images," in *International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, vol. 2, Jan. 2011, pp. 745 –747.

[130] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment using structural distortion measurement," in *IEEE International Conference on Image Processing*, vol. 3, 2002, pp. 5–68.

[131] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, May 2002, pp. 3313–3316.

[132] F. Rahayu, U. Reiter, M. Nielsen, T. Ebrahimi, P. Svensson, and A. Perkis, "Analysis of SSIM performance for digital cinema applications," in *International Workshop on Quality of Multimedia Experience*, Jul. 2009, pp. 23–28.

[133] S. Yasakethu, C. Hewage, W. Fernando, and A. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1969–1976, Nov. 2008.

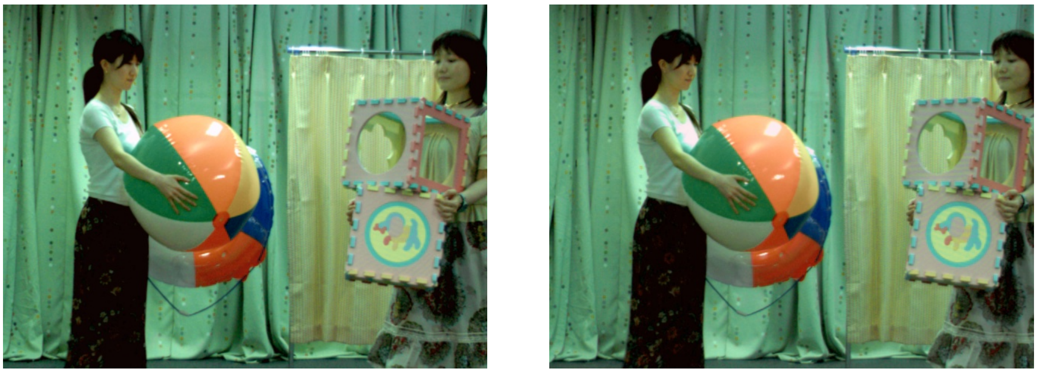
[134] S. Kosov, T. Thormählen, and H.-P. Seidel, "Accurate real-time disparity estimation with variational methods," in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5875, pp. 796–807.

[135] "H.264/AVC reference software with stereo high profile (version 18.2)," Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Feb. 2012. [Online]. Available: http://iphome.hhi.de/suehring/tml

Appendix A

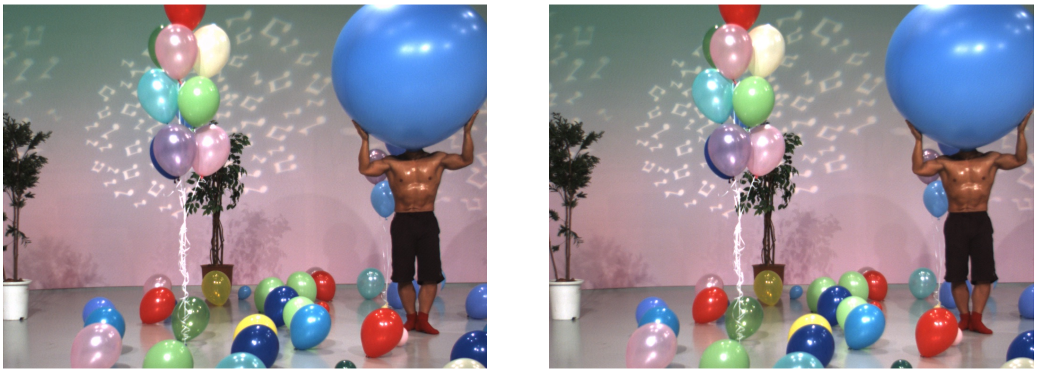
Test sequences

This appendix shows the original test stereoscopic sequences. The sequences Akko & Kayo, Champagne Tower, Balloons, Dog, Kendo and Pantonime were obtained under the permission of Tanimoto Lab at Nagoya University (available at `http://www.tanimoto.nuee.nagoya-u.ac.jp/`).



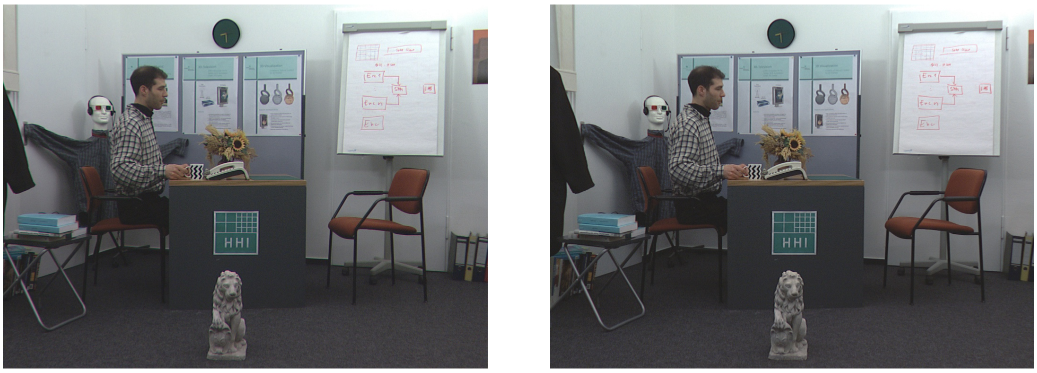
(a) View ID 50 (b) View ID 49

Figure A.1: Stereo sequence Akko & Kayo, 640 × 480.



(a) View ID 3 (b) View ID 2

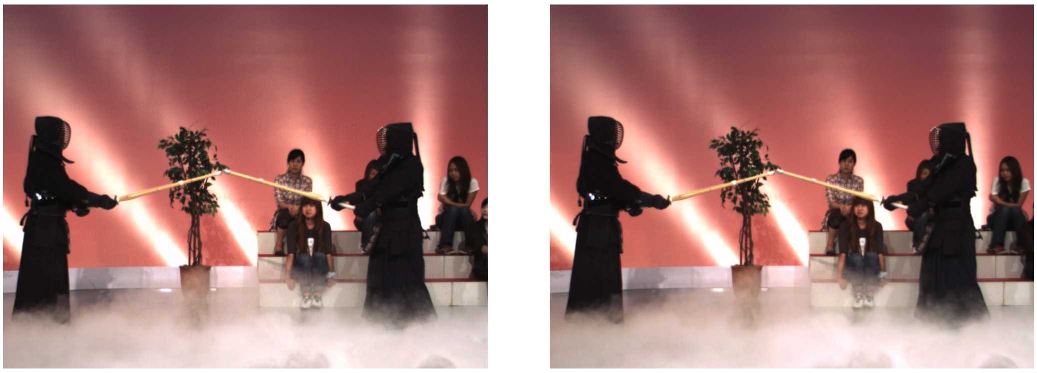
Figure A.2: Stereo sequence Balloons, 1024 × 768.



(a) View ID 9

(b) View ID 8

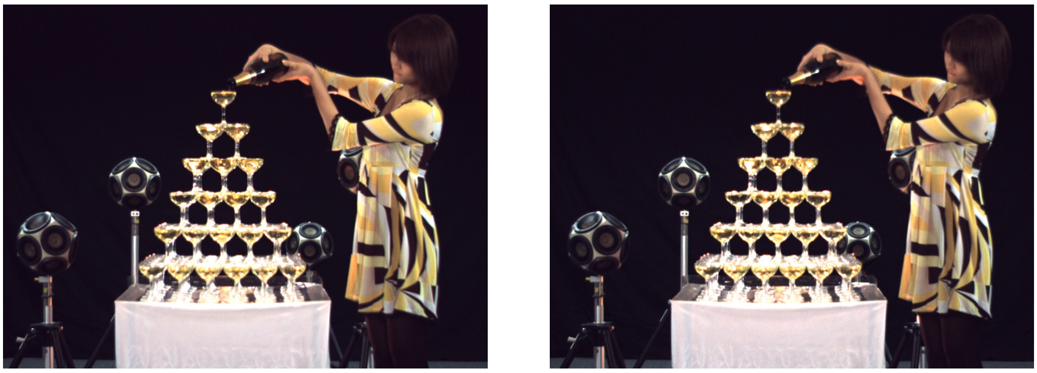
Figure A.3: Stereo sequence Book Arrival, 1024 × 768.



(a) View ID 3

(b) View ID 2

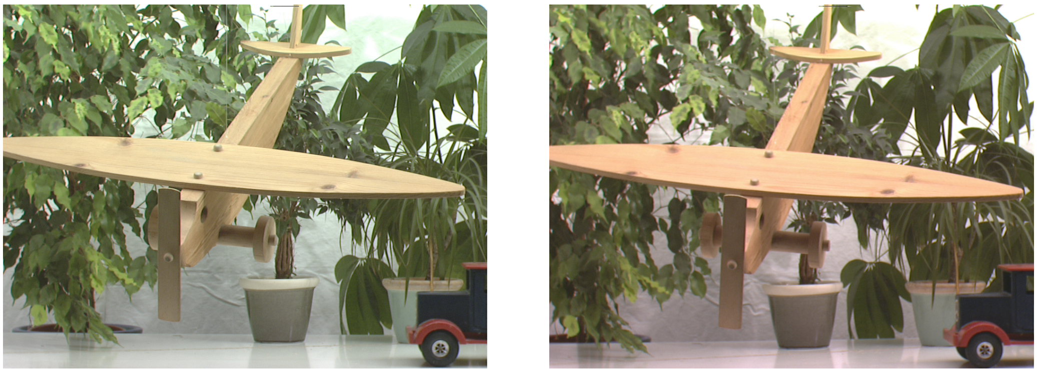
Figure A.4: Stereo sequence Kendo, 1024 × 768.



(a) View ID 40

(b) View ID 39

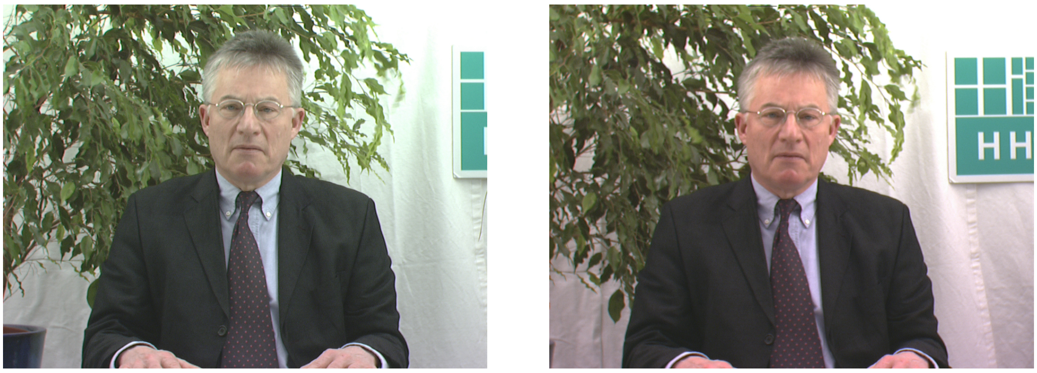
Figure A.5: Stereo sequence Champagne Tower, 1280 × 960.



(a) View ID 3

(b) View ID 4

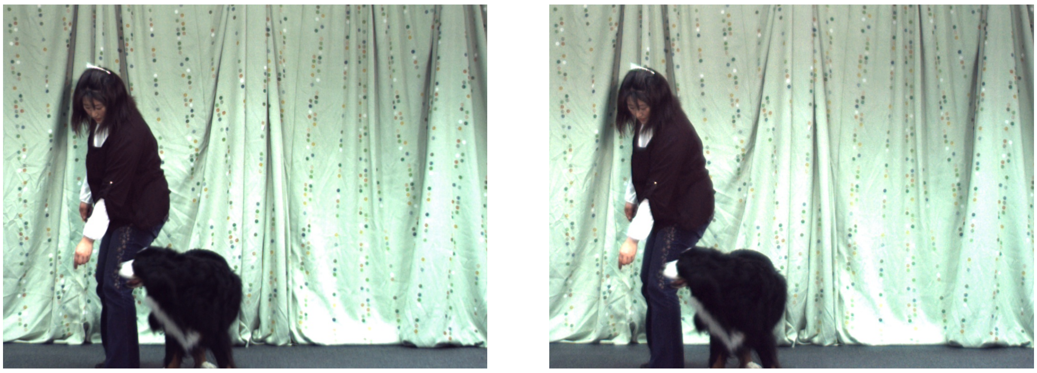
Figure A.6: Stereo sequence Jungle, 1024 × 768.



(a) View ID 3

(b) View ID 4

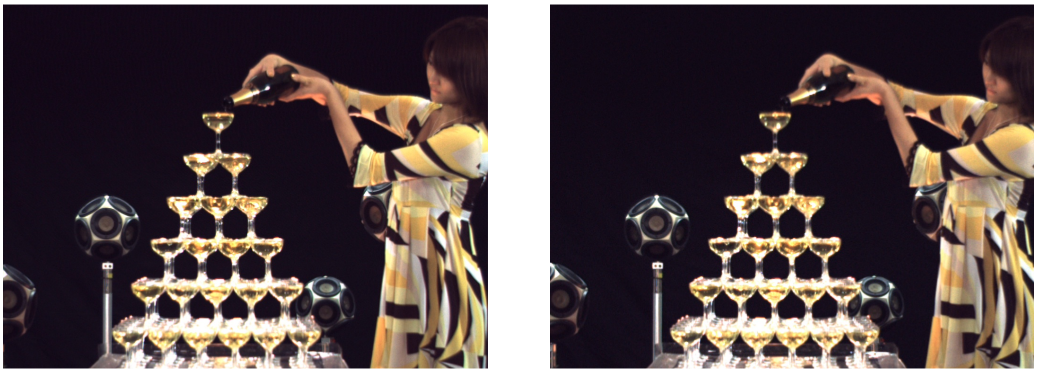
Figure A.7: Stereo sequence Uli, 1024 × 768.



(a) View ID 40

(b) View ID 41

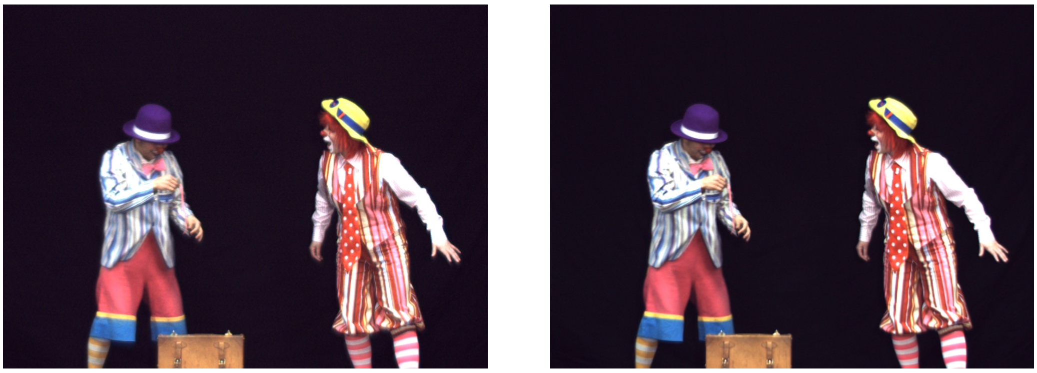
Figure A.8: Cropped version of the stereo sequence Dog, 1024 × 768.



(a) View ID 40

(b) View ID 39

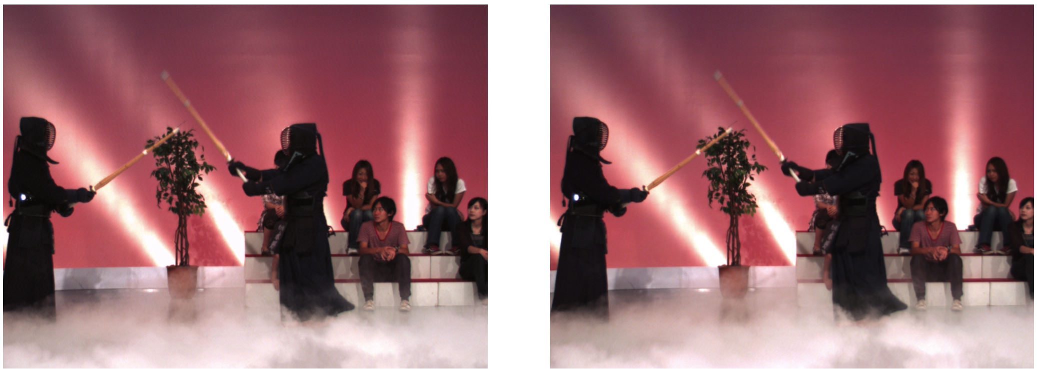
Figure A.9: Cropped version of the stereo sequence Champagne Tower, 1024 × 768.



(a) View ID 2

(b) View ID 3

Figure A.10: Cropped version of the stereo sequence Pantonime, 1024 × 768.



(a) View ID 3

(b) View ID 2

Figure A.11: Frame 18 of the sequence Kendo used to assess the performance of the proposed concealment methods.

Appendix B

Published paper

This appendix presents the published paper, resulted from the research work done during this dissertation.

- J. Carreira, P. Assunção, N. Rodrigues, S. Faria, ”Frame Loss Concealment for 3D video Decoders Based on Disparity-Compensated Motion Field”, *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Zurich, Switzerland, October 2012.