Coding for improved perceived quality of 2D and 3D video over heterogeneous networks.

Linda S. Karlsson

Mittuniversitetet
MID SWEDEN UNIVERSITY

Department of Information Technology and Media
Mid Sweden University

Thesis No. 87
Sundsvall, Sweden
2010

Mittuniversitetet
Informationsteknologi och medier
SE-851 70 Sundsvall
SWEDEN

ISBN 978-91-86073-78-7
ISSN 1652-893X

Akademisk avhandling som med tillstånd av Mittuniversitetet framlägges till offentlig granskning för avläggande av teknologie doktorsexamen. Tid och plats för disputation torsdagen den 27 maj 2007, 10.00 - 13.00, i O111, Mittuniversitetet, Holmgatan 10, Sundsvall.

©Linda S. Karlsson, april 2010

Tryck: Tryckeriet Mittuniversitetet

To My Family And Friends

Abstract

The rapid development of video applications for TV, the internet and mobile phones is being taken one step further in 2010 with the introduction of stereo 3D TV. The 3D experience can be further improved using multiple views in the visualization. The transmission of 2D and 3D video at a sufficiently perceived quality is a challenge considering the diversity in content, the resources of the network and the end-users. Two problems are addressed in this thesis. Firstly, how to improve the perceived quality for an application with a limited bit rate. Secondly, how to ensure the best perceived quality for all end-users in a heterogeneous network.

A solution to the first problem is region-of-interest (ROI) video coding, which adapts the coding to provide a better quality in regions of interest to the viewer. A spatio-temporal filter is proposed to provide codec and standard independent ROI video coding. The filter reduces the number of bits necessary to encode the background and successfully re-allocate these bits to the ROI. The temporal part of the filter reduces the complexity compared to only using a spatial filter. Adaption to the requirements of the transmission channel is possible by controlling the standard deviation of the filter. The filter has also been successfully applied to 3D video in the form of 2D-plus-depth, where the depth data was used in the detection of the ROI.

The second problem can be solved by providing a video sequence that has the best overall quality. Hence, the best quality for each part of the network and for each 2D and 3D visualization system over time. Scalable video coding enables the extraction of the parts of the data to adapt to the requirements of the network and the end-user. A scheme is proposed in this thesis that provides scalability in the depth and view domain of multi-view plus depth video. The data are divided into enhancement layers depending on the content's distance to the camera. Schemes to divide the data into layers within a view and between adjacent views have been analysed. The quality evaluation indicates that the position of the layers in depth as well as the number of layers should be determined by analysing the depth distribution. The front-most layers in adjacent views should be given priority over the others unless the application requires a high quality of the center views.

Sammanfattning

Den snabba utvecklingen av videoapplikation för TV, Internet och mobiltelefoner tar ytterliggare ett steg i och med introduceringen av stereo 3D TV under 2010. Upplevelsen av 3D kan förstärkas ytterliggare genom att använda multipla vyer i visualiseringen. Skillnaden i innehåll, nätverksresurser och slutanvändare gör överföring av 2D och 3D video med en tillräcklig hög upplevd kvalitet till en utmaning. För det första, hur man ökar den upplevda kvalitén hos en applikation med en begränsad överföringshastighet. För det andra, hur man tillhandahåller den bästa upplevda kvalitén hos alla slutanvändare i ett heterogent nätverk.

Region-of-interest (ROI) videokodning är en lösning till det första problemet, vilken anpassar kodningen för att ge högre kvalitet i regioner som är intressanta för användaren. Ett spatio-temporalt filter är föreslaget för att tillhandahålla codec- och standardoberoende ROI videokodning. Filtret reducerar antalet bitar som krävs för att koda bakgrunden och omfördelar dessa till ROI:t. Den temporala delen av filtret minskar komplexiteten jämfört med att använda enbart spatiala filter. Filtret kan anpassas till överföringshastigheten genom att ändra standardavvikelsen för filtret. Filtret har också använts på 3D video i formen 2D-plus-depth, där djupdata användes i ROI detektionen.

Det andra problemet kan lösas genom att tillhandahålla en videosekvens som har högsta möjliga kvalitet i hela nätverket. Därmed även den bästa kvaliteten för for varje del av nätverket och för varje 2D- och 3D-skärm. Skalbar videokodning gör det möjligt att extrahera delar av datan för anpassning till de rådande förutsättningarna. En metod som ger skalbarhet i djupet och mellan kameravyer hos multi-view plus depth video har föreslagits. Videosekvensen delas upp i lager beroende på innehållets avstånd till kameran. Metoder för att fördela data över lager i djupet och mellan närliggande vyer har analyserats. Kvalitetsutvärderingen visar att lagrens position i djupet och antalet lager bör bestämmas utifrån fördelningen av djupdata. De främsta lagren i närliggande vyer bör ges högre prioritet om inte applikationen kräver hög kvalitet hos vyer i centrum.

Acknowledgements

Firstly, I would like to thank my supervisors Docent Mårten Sjöström and Professor Youzhi Xu for providing the necessary help and guidance making this thesis possible. In particular, Docent Mårten Sjöström for all the questions that has helped to make it possible for others to to understand my work. I also thank Professor Tingting Zhang, who was the main supervisor for the first part of the thesis. I would also like thank the others involved in the Realistic 3D and MediaSense groups for their input and discussions. In particular, Dr. Roger Olsson for all the valuable comments and discussion of my work and Professor Theo Kanter for reading and commenting the thesis. I would like to thank Professor Karl W Sandberg, Rolf Dahlin and Dr. Kjell Brunnström for the advice on subjective tests and all participants in the subjective tests.

Many thanks the other members of the Information and Communication System Division at Mid Sweden University for providing a friendly and welcoming environment. All the interesting discussions of various nature during breaks have helped me stay motivated.

I would also like to express my gratitude to the Swedish Graduate School in Telecommunications for providing interesting summer schools, qualified courses and financial support. In addition, I would like to thank the EU Objective 1 and 2-programmes for funding my research.

Last, but far from least, I would like to give my thanks to my family and friends for making sure that my life does not only consist of research and studies. My brother for being my constant companion at concerts and festivals. My parents, Marita and Åke, for their never-ending love and support. My friend Marie for sharing my enthusiasm of video processing and music and my friend Marcas for his support and help to unwind at stressful times.

Sundsvall 2010-04-20

Linda Karlsson

Contents

Abstract **v**

Abstract **vii**

Acknowledgements **ix**

List of Papers **xv**

Terminology **xvii**

1 Introduction **1**

1.1 Two-dimensional video 2

1.2 Three-dimensional video 2

1.3 Perceptual quality 5

1.4 Region-of-interest video coding 6

1.5 Scalable video coding 8

1.6 2D and 3D Video coding standards 8

1.7 Overall aim 9

1.8 Scope 9

1.8.1 ROI video coding 10

1.8.2 Scalable video coding 10

1.9 Concrete and Verifiable Goals 11

1.10 Outline 11

1.11 Contributions 11

2 2D and 3D video coding **15**
2.1 Video source coding 15
2.1.1 Block-based hybrid coding 16
2.2 Multi-view video 18
2.2.1 Multi-view representations 20
2.2.2 Multi-view coding 22
2.2.3 Multi-view plus depth coding 25
2.3 Rendering 26
2.3.1 3D warping 26
2.3.2 Merging of the views 28
2.3.3 View synthesis artifacts 29
2.3.4 Layered approaches to rendering 30
2.4 Video quality 30
2.4.1 Perception of depth 30
2.4.2 Objective quality measures 31
2.4.3 Subjective quality measures 32

3 Region-of-interest coding **39**
3.1 ROI detection 40
3.1.1 Visual attention 40
3.2 Bit allocation 41
3.2.1 Spatial bit allocation 42
3.2.2 Temporal bit allocation 43
3.2.3 Combinations of spatial and temporal bit allocation 43
3.3 3D ROI video coding 44
3.4 Foveated coding 44
3.5 Object-based coding 44
3.6 Contributions 45
3.6.1 Paper I 46
3.6.2 Paper II 46

3.6.3 Paper III 47
3.6.4 Paper V 48

4 Multi-view plus depth scalable video coding 49
4.1 Scalability in 2D 49
4.2 Scalability in 3D 51
4.3 Contributions 54
4.3.1 Paper V 56
4.3.2 Paper VI 56
4.3.3 Paper VII 57

5 Conclusions 59
5.1 ROI video coding 59
5.2 Multi-view plus depth scalable video coding 60
5.3 Future works 61
5.3.1 ROI video coding 61
5.3.2 Multi-view plus depth scalable video coding 62

Bibliography 63

Biography 73

List of papers

This thesis is based on the following papers:

I L. S. Karlsson and M. Sjöström. Improved ROI video coding using variable Gaussian pre-filters and variance in intensity. In *2005 IEEE International conference on Image Processing, ICIP, Genua, Italy*, Vol. 2, pp. 313-316, 2005.

II L. S. Karlsson, M. Sjöström and R. Olsson. Spatio-temporal filter for ROI video coding. In *14th European Signal Processing Conference, EUSIPCO, Florence, Italy*, 2006

III L. S. Karlsson and M. Sjöström. A Spatio-Temporal Filter for Region-of-Interest Video Coding. Submitted to *EURASIP Journal on Image and Video Processing*

IV L. S. Karlsson and M. Sjöström. Region-of-Interest 3D Video Coding based on Depth Images. In *3DTV Conference - True Vision - Capture, Transmission and Display of 3D Video, 3DTV-CON, Istanbul, Turkey*, 2008

V L. S. Karlsson and M. Sjöström. Multiview plus Depth Scalable Coding in the Depth Domain. In *3DTV Conference - True Vision - Capture, Transmission and Display of 3D Video, 3DTV-CON, Potsdam, Germany*, 2009

VI M. Sjöström and L. S. Karlsson. Performance of scalable coding in depth domain. In *SPIE: Conference on Stereoscopic Displays and Applications XXI, San Jose, CA, USA*, 2010

VII L. S. Karlsson and M. Sjöström. Multiview scalable video coding based on depth data distribution. Submitted to *IEEE Transactions on Circuits and Systems for Video Technology*

Terminology

Abbreviations and Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
B	8x8 pixels Block
B-frame	Bidirectionally predictive coded frame
CABAC	Context-based Adaptive Binary Arithmetic Coding
CGS	Corse Grain quality Scalability
DCT	Discrete Cosine Transform
DES	Depth Enhanced Stereo
FGS	Fine Grain quality Scalability
fps	Frames Per Second
GOP	Group Of Pictures
GGOP	Group of GOPs
HVS	Human Visual System
I-frame	Intra coded frame
ITU	International Telecommunication Union
kbps	Kilobits Per Second
MB	16x16 pixels Makroblock
MGS	Medium Grain quality Scalability
MOP	Matrix of pictures
MPEG	Moving Photographic Experts Group
MSE	Mean Square Error
MVC	Multi-View Codin
MVD	Multi-View plus Depth
LDV	Layered Depth Video
P-frame	Predictive coded frame

PSNR	Peak Signal to Noise Ratio
RGB	Red, Gren, Blue.
ROI	Region Of Interest
SAD	The Sum of Absolute Differences.
	SDM Sample Density Measure
SNR	Signal to Noise Ratio
SOM	Self-Organizing Map
SSIM	Structural Similarity Metric
SP	Spatial
SPTP	Spatio-temporal
SVC	Scalable Video Coding
TSL	Tint, Saturation, Luminance
TP	Temporal
VLC	Variable Length Code
VQM	Video Quality Metric
YCbCr	Colorspace with one luma component and two chrominance components, blue and red

Mathematical Notation

α	A linear weight.
B	Blue component of the RGB color space.
C_b	The blue-difference chrominance component in the YCbCr color space.
C_r	The red-difference chrominance component in the YCbCr color space.
$\mathbf{D}$	A 4×4 matrix that contains R and t.
G	Green component of the RGB color space.
$\mathbf{K}$	Calibration matrix of a camera.
$I(u, v)$,	The color values of pixel (u, v).
$I_L(u_L, v_L)$,	The color values of pixel (u_L, v_L).
$I_R(u_R, v_R)$,	The color values of pixel (u_R, v_R).
$\mathbf{M}$	3D point $(x, y, z, 1)$ in the scene (homogenous coordinates).
$\mathbf{m}$	A point in the image plane (u, v) of a camera.
$\mathbf{m_L}$	A point in the image plane (u_L, v_L) in the left view.
$\mathbf{m_v}$	A point in the image plane (u, v) in a virtual view.
MSE	Mean square error.

$\mathbf{N}$	Total number of pixels in a video signal.
$\mathbf{P}$	Projection matrix of a camera.
$\mathbf{P_L}$	Projection matrix of the left view.
$\mathbf{P_v}$	Projection matrix of the virtual view.
$PSNR_{dB}$	Peak signal to noise ratio.
R	Red component of the RGB color space.
$\mathbf{R}$	Rotational matrix of a camera.
r_{ij}	Component in row i, column j of R.
$\mathbf{t}$	Transformation vector of a camera.
$\mathbf{t_L}$	Transformation vector of the left view.
$\mathbf{t_R}$	Transformation vector of the right view.
t_i	Component in row i of t.
(u, v)	Coordinate system of the image plane of a camera.
(u_L, v_L)	Coordinate system of the image plane of the left camera.
(x, y, z)	Coordinate system of a 3D scene.
x_i	Pixel i in the original video sequence.
y_i	Pixel i in the distorted video sequence.
Y	The intensity component in the YCbCr color space.
Z_{far}	The maximum depth value of the scene.
Z_{near}	The minimum depth value of the scene.

Chapter 1

Introduction

The interest in various types of media applications, including movies, TV-programs, small video sequences, video conversations and surveillance applications has created a demand for delivering large amounts of video data. High performance fiber networks make it possible to transmit multiple video sequences at the same time. In reality, wireless networks, mobile networks and low performance internet connections may be used to transmit this data and the applications may address a wide variety of devices. In addition three dimensional video is gaining ground as a form of entertainment, increasing the demand on the networks and the range of displays.

The task of transmitting a two-dimensional (2D) or three-dimensional (3D) video sequence to all of the end-users in this kind of network and providing the highest perceived quality for each end-user is a difficult problem to solve. This thesis addresses possible solutions regarding the use of source coding to improve the overall quality under these limitations. The quality of a source coded video sequence is dependant of the bit-rate. Hence, the source coding must adapt to any bit-rate limitations in the network to enable a transmission with an as high perceived quality as possible. In this thesis a heterogeneous network is defined as a network, where a variation in the bit-rate budget exists both locally and over time. Two possible solutions to the problem are proposed.

1. In low bit rate video transmission the necessary encoding causes reduced quality in regions of the video sequence with high detail and motion content. The perceived quality is experienced as particularly poor if the region contains information that is important to the viewer. For example, in a video-conferencing sequence the face is of interest. Region of interest (ROI) coding uses this information to increase the quality in regions that are of interest to the viewer at the expense of the quality in the background.

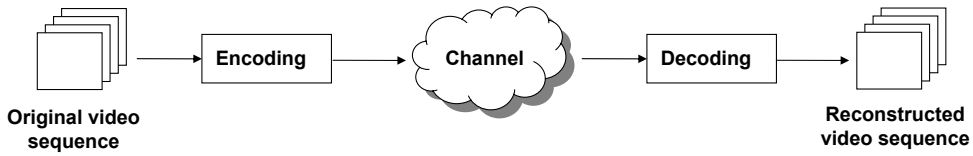


Figure 1.1: Video transmission system

2. The perceived quality and bit rate can be adapted to the conditions of the receiver using scalable video coding (SVC). This enables the extraction of parts of one transmitted video sequence. This is useful when the full video sequence either contains more information than the receiver requires when displaying the video, or it exceeds the bit rate budget. An example is 3D TV, which is transmitted over a network of various bit rate requirements and where the end-users are equipped with either 2D or 3D displays. The various 3D displays use different techniques to depict 3D video. SVC enables the extraction of the exact 3D video data required for each display in the network and the best quality under the bit rate budget of each end-user. Thus, the overall perceived quality in the network can be increased without the cost of sending one video sequence per additional requirement.

1.1 Two-dimensional video

A digital video sequence consists of a sequence of images, which are called frames. These are extracted at a sufficiently small time interval to preserve the continuity in the sequence. Each frame is built up from small picture elements, pixels, which describe the color at that point in the frame using three different components. A large amount of data is needed to completely describe the video sequence. The video sequence is therefore encoded to reduce the amount of data. This makes transmission over channels with limited bandwidth possible. An example of a video transmission system can be found in fig. 1.1, where the sequence is compressed by the encoder before it is transmitted over the channel. In the decoding step the sequence is reconstructed. The reconstructed video sequence contains errors introduced both by the removed information in the compression and by the distortion in the channel.

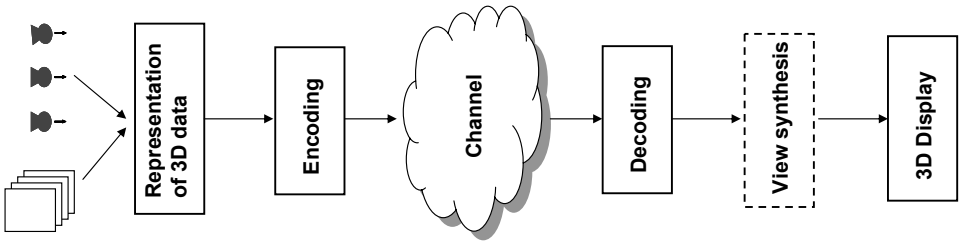


Figure 1.2: 3D video transmission system

1.2 Three-dimensional video

A 3D experience can be provided in a variety of ways, such as multi-view, holography and volumetric methods. The 3D data can be stored and transmitted in formats that range from a set of 2D video sequences of a scene (multi-view), captured at different camera positions, to a graphical description of the scene. In this thesis the focus is on 3D video of a real scene. The properties of such a scene can be described by multiple views of the scene of a sufficiently high number. The captured views in a multi-view contain the high level of details found in a natural scene that are difficult to represent using geometric models.

The additional dimension that the 3D video provides increases the complexity of the whole chain from the 3D content creation to displaying the 3D video (See figure 1.2.), where the choice in all parts has an influence on the others.

- The 3D content is either created by capturing the scene with a set of cameras, or converting existing 2D video into 3D video. The choice of the type, number and position of cameras is connected to the other parts of the system.
- Several representations of multi-view video exists and an overview can be found in section 2.2.1. The transmission of all views requires a high bit-rate in general and is not necessary. Fewer views can be transmitted if depth data is included for each view (multi-view plus depth (MVD)) or by using several layers of data to describe what is occluded (layered depth video (LDV) and depth enhanced stereo (DES)).
- The encoding and decoding of multi-view sequences is based on the coding of 2D video, the difference being that redundant information between the camera views can be used to further compress the data. Depth data of an MVD sequence can be encoded as multi-view, but could be encoded more efficiently if its characteristics are fully considered. The specific characteristics of the occlusion layers in LDV and DES makes efficient coding using the current standards

difficult. More details can be found in section 2.2.2.

- The MVD, LDV and DES representations may not contain the exact views that are needed to display 3D. The missing views are then synthesized based on the available data. An overview of view synthesis algorithms for MVD can be found in section 2.3.

- There are several techniques used to display 3D video. These techniques depict stereo or multi-view video that requires additional hardware and functionalities in the form of glasses and head tracking or have built in optics that create the illusion of 3D [1]. The transmitted data should also be able to be displayed directly on to a 2D screen, when the receiver does not have a 3D display or a limited bit-rate.

1.3 Perceptual quality

Video source coding in general aims to preserve quality, while reducing the bit rate. In most cases the quality is defined by the extent of the error introduced by the compression independent of the error's position in the video sequence. This is a simplification, which disregards the complexity of the human visual system (HVS). The perceptual quality is highly dependent on the information being transmitted at the location of the error. In regions containing details and particularly important semantic content, such as the face in video conferencing, the impact of an error is much greater than in the less important background. In 3D video the position in depth and the error's influence on the depth cues might also play an important role.

The perceived quality of 3D video depends on several factors, including the representation format, the compression, and the synthesis of views that are not transmitted directly. Another factor is the visualization system used to display the 3D video to the viewer. Hence, measuring the perceived quality in a separate part of the system, such as the compression, is a challenge.

The most reliable metric for perceived quality is an extensive subjective test, however such a test is time-consuming with high requirements on the setup. Hence, a smaller subjective test, visual evaluation by a few specialists or objective metrics that give information of quality in parts of the 2D or 3D video sequence can provide an indication of the perceived quality. More about methods for detecting quality in both 2D and 3D are presented in section 2.4.

1.4 Region-of-interest video coding

Video coding is preformed to optimize the average quality under the constraints of a limited bit-rate. The result of the compression is reduced quality in regions of the sequence containing high detail and motion content, such as a talking head. An example is depicted in fig 1.3.a), containing an image from a video sequence. The face is moving compared to the background and contains details. Hence, the perceived quality of this video sequence will be reduced, since the errors are concentrated to the facial region.

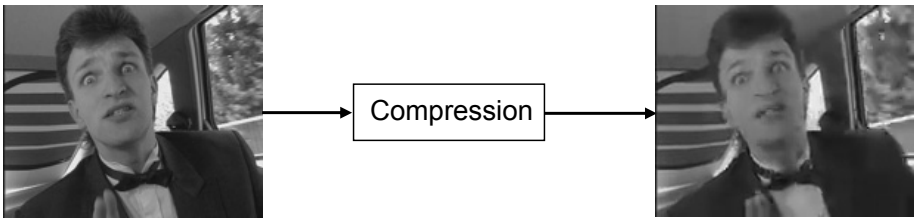
If an interesting region can be detected, ROI video coding can be applied. ROI coding increases the quality in the interesting region at the expense of the quality in the background as in fig 1.3.b. The result is an increase in perceived quality without increasing the bit rate.

The ROI video coding consists of two main steps. The ROI must firstly be detected, which requires previous knowledge of what a human would find interesting in the sequence; the perceived quality may even be reduced if the correct ROI is *not* detected. Secondly, the regions of the video that are not classified as ROI (background) are compressed more strongly than the ROI. This is achieved by bit allocation, which controls how many bits will be allocated to the different parts of the video sequence. More details concerning ROI video coding are presented in chapter 3.

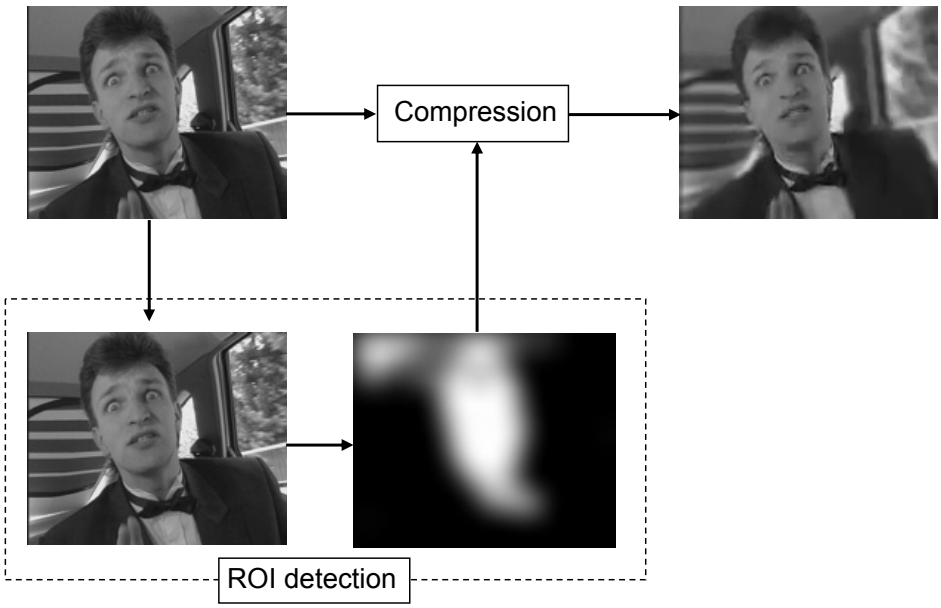
1.5 Scalable video coding

Scalable video coding (SVC) makes it possible to use one video sequence for the entire heterogenous network and to extract the necessary information from this sequence depending on the characteristics of a specific receiver. This partitioning of the data in layers is based on how important that data is for the perceived quality of the video sequence. The first layer, the base layer, must then contain the necessary data to play out the video sequence in the least acceptable quality. The rest of the data is partitioned into enhancement layers that can be added one by one to increase the quality of the video sequence. If all layers are used the quality is approximately the same as when traditional video coding is applied to that particular sequence with the same level of compression.

In most cases the resulting bit stream is reordered to facilitate an easy extraction of the parts that can be transmitted in that part of the network or that a receiver requires. An example is shown in figure 1.4. The bit stream of encoded video sequence is in general transmitted frame by frame as in the top of the figure. Scalability can



(a) Video Coding



(b) ROI Video Coding

Figure 1.3: ROI video coding. a) All parts of the frame are compressed equally, when ordinary video coding is applied. b) The perceived quality can be improved by applying more compression to the background than the ROI.

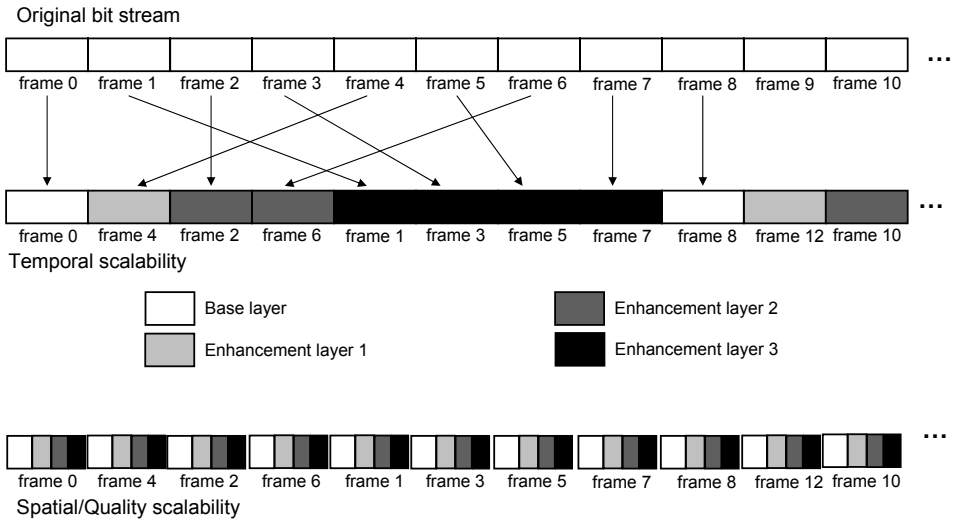


Figure 1.4: Examples of a bit stream. The bit stream produced by traditional encoding (top), temporal SVC (middle), and spatial/quality SVC (bottom). The decoder can then extract just the base layer or the number of enhancement layers that it can afford to increase the quality of the video sequence.

be achieved in various ways and one of them, temporal scalability, is depicted in the middle of figure 1.4. Temporal scalability provides the means to extract frames at the desired frame rate. SVC can also be used to provide frames at various resolutions (spatial scalability) or with various amounts of details (quality scalability). An example of how the bit stream can be organized for spatial or quality scalability is found at the bottom of figure 1.4. In addition to reordering the data, the encoding of the data must be adapted to the characteristics of the data in the layers. If the encoding of the data of a layer depends on data from higher layers, then errors occur if the higher layers are not available. These types of errors might cause errors in several frames in a row. More details on SVC can be found in chapter 4.

1.6 2D and 3D Video coding standards

The Moving Pictures Reference Group MPEG [2], which is a working group of ISO/IEC and the International Telecommunications Union ITU [3] are responsible for producing the main standards available for video compression standards. These include MPEG-2 [4], H.261 [5] and H.263 [6], which are intended for different applications with block-based hybrid coding as the common factor. (See section 2.1.1). The standard MPEG-4 [7] includes additional features such as object-based coding. (See sec-

tion 3.5). The most recent standard is H.264 [8][9], also called MPEG-4 AVC, which has a higher compression efficiency than previous standards.

Two extensions in H.264 are of interest in this thesis. The MPEG-C part 3 (ISO/IEC 23002-3) [10] supports coding of the multi-view and 2D plus depth representations. Scalability approaches such as temporal, spatial and quality scalability are supported by the SVC extension of H.264 [11].

These standards describe the features supported by the encoder, define the syntax and semantics of the bit stream and the manner in which the transmitted bit stream should be parsed and decoded [12]. This enables some freedom when implementing the various codecs.

1.7 Overall aim

The primary goal of this thesis is to analyze how source coding can be used to increase the perceived quality of all end-users in a heterogeneous network. The network is heterogeneous in the sense that the limitations on bit rate may vary locally and over time. The display systems used by the end-users may also vary. The knowledge that regions with certain characteristics are more important to perceived quality than others is considered in this thesis.

1.8 Scope

This thesis will address source coding of 2D and 3D video sequences captured from real scenes. View synthesis is included in the thesis as a tool to provide the necessary views to display 3D from a coded 3D video sequence. The operations performed on the encoded data and its transmission are not considered in this thesis, except that parts of the network are assumed to be subjected to a limited bit-rate. No transmission errors are taken into account.

Perceived quality is difficult to measure without resorting to an extensive subjective analysis that requires both time and careful planning of the execution of the tests. Due to the limitations of time and other resources objective quality metrics based on the peak-signal-to-noise-ratio (PSNR) were used in the quality evaluation, since the PSNR metric is still used extensively in the field of video coding. Extra information about the perceived quality was provided by PSNR metrics for content at either specific regions (ROI and background) or in the 3D case for each rendered view and in relation to the depth of the scene. The results was either verified using a minor subjective test or by visual examination of parts of the data by experts. (More

about quality metrics is found in section 2.4.)

1.8.1 ROI video coding

The first part of the thesis deals with how to use ROI video coding to improve perceived quality at a limited bit rate. It is assumed that the ROI has already been successfully detected. Only pre-processing methods are addressed, since methods associated with altering the encoder means that the implementation must be remade each time the encoder is changed. Pre-processing methods in both the spatial and temporal domain are investigated.

In the tests of the spatio-temporal filter only 2D and 2D plus depth video sequences involving communicating humans are considered in order to limit the impact of false detections of the ROI. Thus only sequences were applied, where the face detection algorithm was successful. However, it can be assumed that the methods could be applied to any type of ROI as long as the ROI detection is successful.

The depth data was used in the ROI detection, when the ROI coding approach was extended to 2D plus depth video sequences.

The performance of the spatio-temporal filter was evaluated by a qualitative analysis of the effect the spatial and temporal parts of the filter have on the encoding and on computational complexity. In addition, quantitative tests are performed using objective quality metrics that analyse the ROI region and the background separately. A minor subjective test is used to verify the results.

1.8.2 Scalable video coding

In the second part of the thesis, methods to provide scalable video coding for multi-view data are analyzed. The multi-view plus depth representation was chosen due to its ability to provide quality for data at higher disparity than 2D plus depth at the same time as the cost in bit rate is lower than transmitting all views. Layered depth video (LDV) could decrease the bit rate further, but is not chosen due to the increased coding complexity, higher sensitivity to errors in the depth data and risk of artifacts in synthesized views. The details concerning 2D plus depth and LDV are found in section 2.2.1.

In this thesis the focus is on scalability schemes that utilize the 3D characteristics of the sequence, such as the additional views and the depth information. Therefore, temporal, spatial and quality scalability and other 2D video based scalability methods fall outside of the scope of this thesis.

The focus of the thesis is on the compression and therefore a basic rendering

algorithm will be chosen to make performance tests possible. The rendered sequence is analysed is evaluated using a objective metric to measure the effect the addition of each layer has on the overall quality, as well as on the bit rate.

Subjective tests of 3D video are considered to fall outside of the scope of this thesis, due to lack of time and resources. An indication of the perceived quality of the rendered 3D video sequence is provided by quantitative evaluation. The evaluation uses objective metrics that provide information of the quality at different depths or per view. This is combined with a visual evaluation of selected views.

1.9 Concrete and Verifiable Goals

1. Propose a codec independent region-of-interest approach that is applied in both the temporal and spatial domain to increase the perceived quality in a 2D video sequence at a fixed bit rate.
2. Provide a scheme that extracts regions-of-interest in 2D plus depth video sequences.
3. Propose a scalable coding algorithm of multi-view plus depth data that addresses the depth domain using the concept that some regions in the video sequence are more important than others.

1.10 Outline

The thesis is organized as follows: In chapter 2 an overview of the related works in video coding, multi-view 3D video, rendering and quality measures are presented. A more detailed description of the fields *ROI video coding* and *scalable video coding* are found in Chapters 3 and 4. Each of the chapters contains a presentation of the related works of that specific field and at the end of each chapter the author's contributions to that field are summarized. A summary of the conclusions and future works are given in chapter 5. The eight papers that contain the original contributions can be found in the appendix.

1.11 Contributions

The author's contributions are presented in the seven papers included in this thesis. The author is responsible for the concept and ideas, evaluation criteria, tests, analysis and presentation in all papers. The co-authors have contributed to the formulation

of evaluation criteria and the analysis of the results in all papers. In paper VI the co-author (first author) is also responsible for the main part of the presentation. The main contributions in the thesis are:

1. a codec-independent spatio-temporal filter with a gradual transition from background to ROI to avoid artifacts at the ROI border is proposed in paper II. The filter increases the coding efficiency in both the spatial and temporal domain and has a lower complexity than when only applying the spatial filter proposed in paper I.

2. a theoretical analysis of the effects of the spatial and temporal parts of the spatio-temporal filter concerning coding efficiency, ability to reallocate bits from background and computational complexity (paper II and III). The impact of the choice of filter parameters of the spatial filter on rate-distortion is also analyzed in paper III.

3. an analysis of the objective performance (paper II and paper III) and the spatio-temporal filter's effect on subjective quality (paper III).

4. an approach to detect ROI in a 2D plus depth video sequence using the depth data and additional ROI characteristics and apply the spatio-temporal filter. (Paper IV.)

5. a scalable coding algorithm of the color part of multi-view plus depth data that allows parts of a view to be extracted depending on the depth values in the depth data is proposed in paper V. This is combined with view scalability to ensure the presence of color and depth data at all depths and to increase the quality of the center views.

6. three different algorithms regarding how to extract the layers based on depth data are proposed in paper VI and to use the same scalability approach to the depth data is proposed in paper VII.

7. strategies regarding how to prioritize between the data of the views in the enhancement layers if the multi-view plus depth sequence contains more than three views (paper VII).

8. an analysis of the effect of the layer assignment schemes on rate-distortion (paper VI and VII) and how the quality is affected per view and in relation to depth (paper VII).

Chapter 2

2D and 3D video coding

The vast field of video coding has been thoroughly researched, but the changing methods of communication, new applications and the increased interest in 3D video has raised new research issues. In this thesis the problem of providing high quality 2D and 3D video over heterogeneous networks is addressed, where the bit-rate requirements as well as the types of end devices are variable. The aim of this chapter is to provide the basic knowledge of video coding, 3D in the form of multi-view and the rendering of virtual views from a multi-view plus depth video sequence, necessary to understand the rest of the thesis. The chapter is concluded by an overview of quality measurements.

2.1 Video source coding

A digital 2D video sequence consists of a sequence of frames, in which each frame contains a large number of pixels. The video sequence is perceived as continuous when the pixels are sufficiently small and the frames are densely spaced in time. The perceived quality of the video sequence depends on the number of pixels (resolution) and the number of frames per second (frame rate). At the same time, a video sequence has a predictable appearance, with large areas containing data that are similar both within that frame and compared to the previous frames. This redundancy in the data is the key to compressing the video sequence with as little impact on quality as possible.

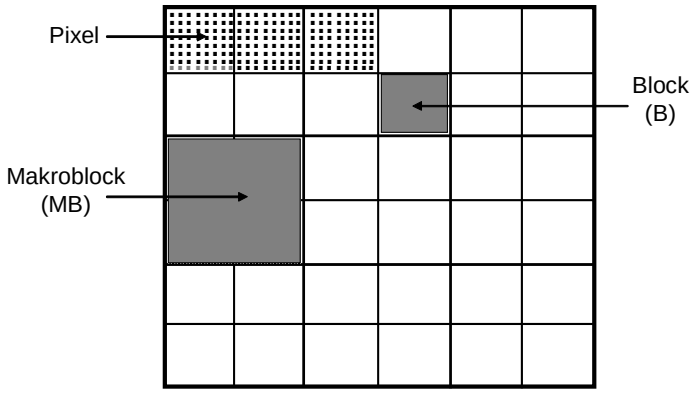


Figure 2.1: The division of a video frame into macro-blocks (MB), blocks (B) and pixels.

2.1.1 Block-based hybrid coding

Most video coding standards are based on block-based hybrid coding. The basic scheme partitions each frame into 16x16 pixel macro-blocks (MB) and these MBs are further partitioned into blocks (B) of 8x8 pixels (See figure 2.1). Each frame is either intra-coded using only information within the frame, or inter-coded, which indicates that information from other frames is also used. These two methods provide the scheme's hybrid character.

In an *intra-coded* frame each B is subject to transform coding using the Discrete Cosine Transform (DCT) (See fig 2.2). The DCT represents the information in each block using a mean value of the block and the deviation from this mean value using combinations of 63 different patterns. Video usually contains several one colored areas, which in this case would be represented as one value instead of 64 values. In the case of deviations from this mean value, only a few of the 63 patterns are generally necessary to express this deviation. Therefore much less information is required to be transmitted with this representation than if the pixel values were transmitted directly. The DCT components are firstly quantized and then encoded using a variable length code (VLC), before they are transmitted.

In the *inter-coded* frames, the fact that only minor changes occur between two frames is used in the encoding. This change can be represented by *motion vectors* that represent the translation of a macro-block compared to its best match in the previous

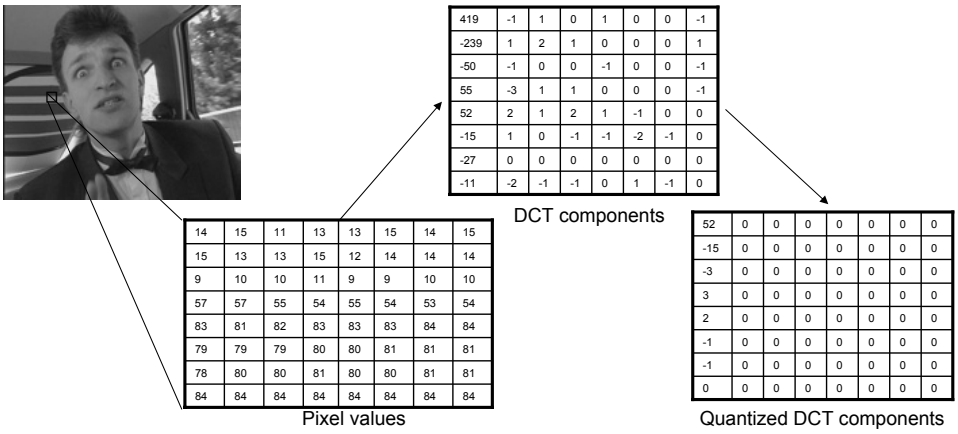


Figure 2.2: Discrete Cosine Transform (DCT) of a block within a frame followed by quantization of the resulting components.

frame. (See fig 2.3). The error between the MB and its best match in the previous frame, the prediction error, is transformed using the DCT and thereafter quantized. The motion vectors and the quantized prediction error are encoded using VLC. The result is transmitted in a bit stream. The bit stream is decoded at the receiver by performing the reverse operations associated with the encoding of both the intra- and inter-coded frames in order to reconstruct the sequence at the receiver side.

The video sequence is divided into group of pictures (GOP) that provide both a point of access and errors in the prediction do not propagate outside of the GOP. (See fig. 2.4.) Three types of frames are in general used in the encoding. I-frames are encoded using strictly intra-coding. Hence, an I-frame provides a point of access and prevents propagation of errors. A P-frame, on the other hand, is encoded using inter-coding with prediction from previous frames and can therefore be compressed more efficiently than an I-frame. The third type of frames, B-frames use both the previous and the next frame in the inter-coding. A combination of these types of frames is in general used to provide the least bit rate with an as little effect on the quality.

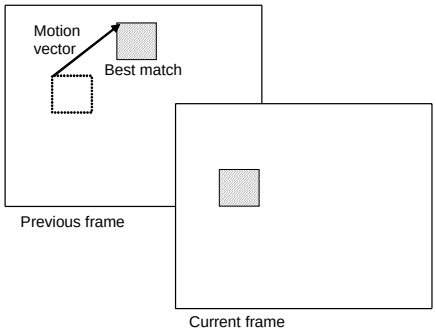


Figure 2.3: The best match from the previous frame is estimated and the motion vector to the position of the best match determined.

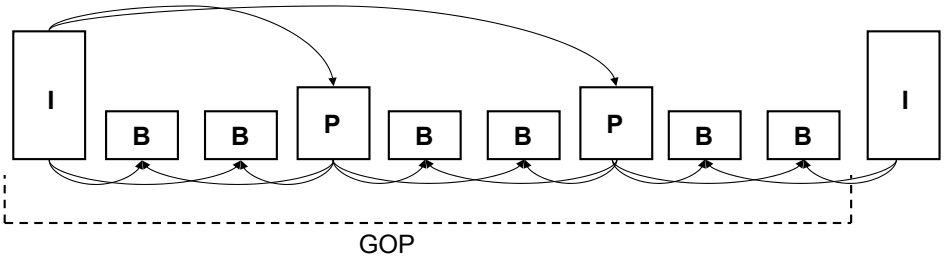


Figure 2.4: A video sequence consists of several group of pictures (GOP) that begin with an I-frame and contains several B- and/or P-frames.

2.2 Multi-view video

The recent interest in 3D technology has spawned a number of different applications including the 3D cinema [13], 3DTV [14], mobile phones [15] and visualization of data in medicine and the industry among others. Techniques exist for viewing movies and games using stereo technology on a PC. In general, additional glasses are required to enable the 3D experience [16]. Several companies have indicated that they will provide solutions for stereo 3D TV concerning both content and equipment in 2010 [17].

Stereo video gives a 3D experience through anaglyphic, polarized or shutter glasses. A problem associated with stereo is that it only provides 3D from one direction, whereas our visual system expects to see different parts of objects if the head is moved. Multi-view can provide all the necessary depth cues [1] and is therefore considered one of the most promising techniques to provide a 3D experience for multiple viewers without the discomforting glasses and with less restrictions on head movement. Several full resolution video sequences are required to depict multi-view. One per view is necessary if the desired views are far apart. Hence, a huge amount of data must be transmitted. Real-time transmission of multi-view in heterogeneous networks is possible if the redundancy between the views is exploited using video coding.

2.2.1 Multi-view representations

The decision with regards how to represent a 3D scene has a high influence on all other parts of the 3D video system. This decision sets the requirements on the capturing process and the encoding in addition to the rendering algorithms. These depend on the data characteristics and what information is available to render the final 3D video sequence for a particular display. The best choice considering complexity, additional processing, bit rate and quality also depends on the content of the scene and the requirements on the number and position of the views that are displayed. The high level of details in the video sequences of a natural scene and the lack of previous knowledge of the scene geometry makes image-based approaches appropriate for natural scenes. [18]

The currently available representations for multi-view video of natural scenes [19] range from transmitting all the data that a display requires (multi-view) to only transmitting the color and depth data of one view (2D-plus-depth). (See top of figure 2.5.a-b.) The latter contains only one view and depth information. Each depth map provides the depth value per pixel of one view represented by the corresponding 2D video sequence. The depth values range between minimum $Z_{near} = 255$ and the

maximum $Z_{far} = 0$, representing the distance to the camera.

Transmitting all views requires a high bit rate, since the bit-rate increases linearly with the number of views. Therefore high quality can be maintained per view independently of the disparity between the views. In the case of a small disparity between the views, the necessary views could be synthesized from the 2D plus depth representation. Transmitting one view and depth information instead of all the views reduces the bit rate substantially. However, a large disparity between the existing data and the views to be synthesized increases the number of artifacts due to disoccluded parts of the scene. Disocclusion occurs when data, occluded in the transmitted view, are supposed to appear in the view to be rendered.

A solution to the problem associated with high bit-rates when transmitting all views and the occlusion problem of 2D plus depth is to either include depth data or occlusion layers to the sequence. Multi-view plus depth (MVD) contains multiple views with depth information for each view (see figure 2.5.c)), whereas layered depth video (LDV) contains the data of one view and the data of occluded parts of that view. Hence, the extra information enables a correct rendering of disoccluded objects. The new approach depth enhanced stereo (DES) complies with the trend in industry to provide stereo video. It enhances the stereo pair by providing additional depth and occlusion layers to extend the adaptability of the representation.

The addition of depth data in the representation increases the ability to create new views from the transmitted data. However, the creation of depth data is still a problem due to the lack of automatic capturing systems that provide accurate and reliable depth maps.

2.2.2 Multi-view coding

The multi-view video sequence is a set of 2D video sequences, one per camera. Therefore, the basics of the 2D video coding can be applied to multi-view sequences using for example the H.264 or MPEG-2 standards. According to Merkle at al in [20], H.264/AVC and hierarchical B-frames have shown to provide the highest coding efficiency. An example of hierarchical B-frames can be found in figure 2.6. The set of 2D video sequences that constitutes a multi-view video sequence contains redundancy with respect to each other. Interview redundancy can be exploited in a similar manner as redundancy between frames using motion compensation [21]. (Motion compensation between frames is described in section 2.1.1.) This is achieved by creating what is called a group of GOP (GGOP) [22] or matrix of pictures (MOP) [23], which describes set of views that are grouped together in a similar manner as the previously described GOP. (Section 2.1.1.) The coding efficiency can then be improved by allowing interview prediction as well as interframe prediction (See figure 2.7).

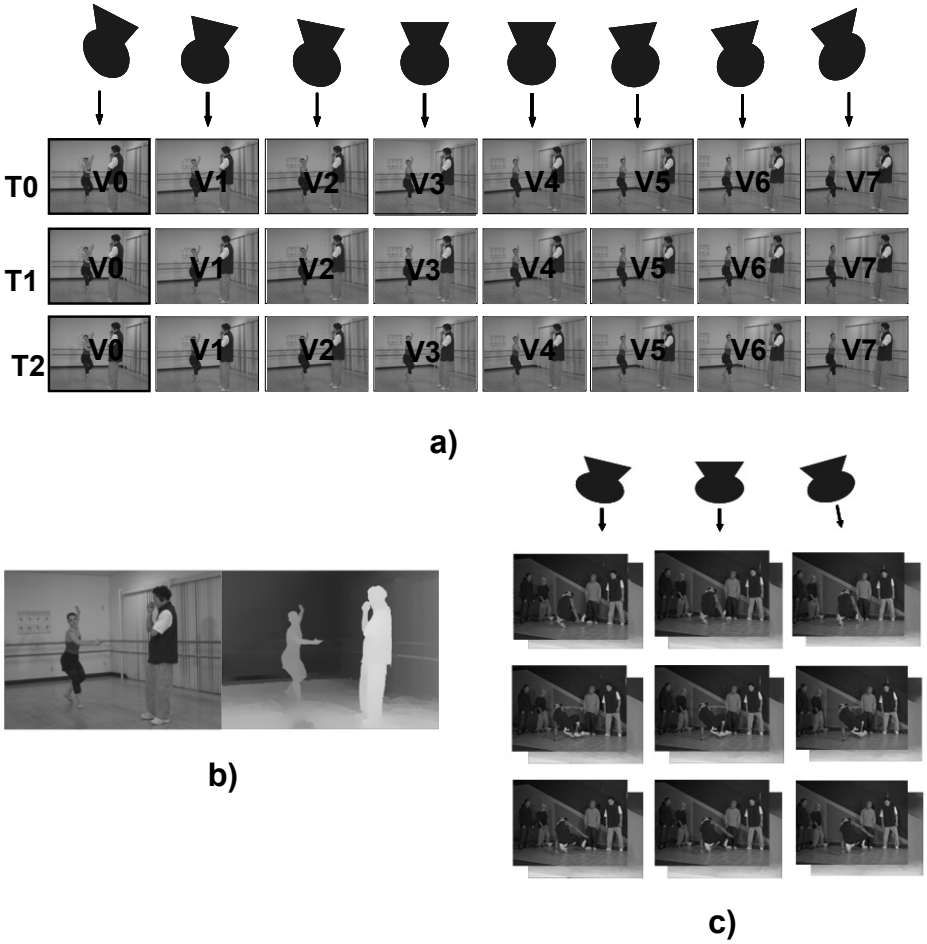


Figure 2.5: Examples of multi-view representations include a) transmitting all views, b) 2D plus depth and c) multi-view plus depth.

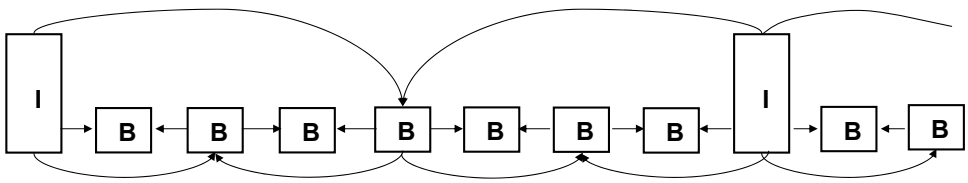


Figure 2.6: An example of hierarchical B-frames.

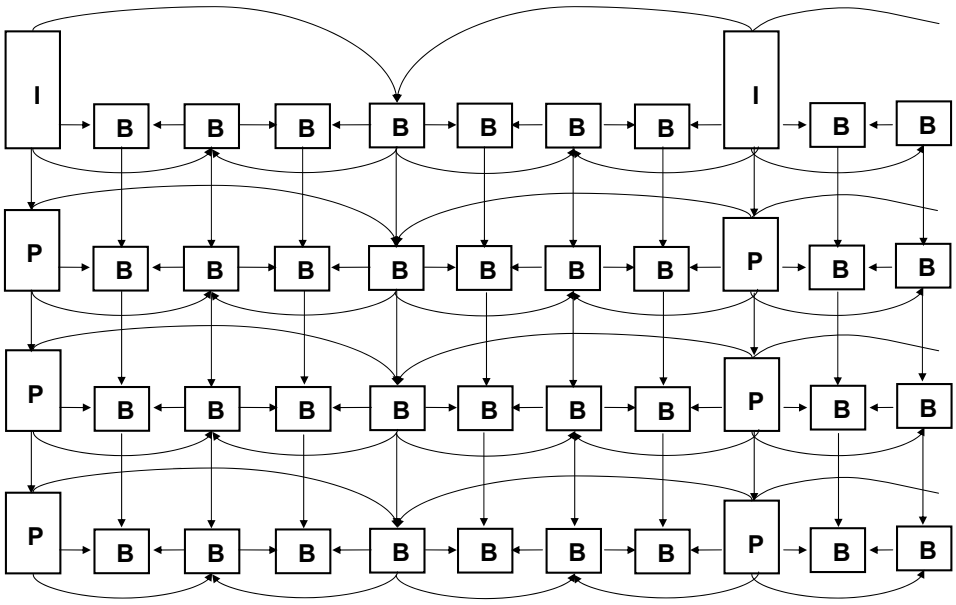


Figure 2.7: A four view multi-view sequence encoded using hierarchical B-frames and inter-view prediction.

One problem associated with the interview prediction in figure 2.7 is the increased complexity. One option is to only allow interview prediction at key frames. Interview prediction at key frames reduces the coding efficiency slightly compared to its use for all frames. However, the results in [20] indicate that the reduction in complexity of the encoder is substantial. In addition, the interview prediction might not have any effect on the coding efficiency in the case of sparsely positioned cameras in the capturing process. Disparity compensation techniques that are based on the characteristics of interview redundancy can be used to further improve the coding efficiency. [24] Another problem is the lighting in the scene and the calibration of the cameras at the moment of capture. The lighting and color of objects in the scene vary depending on the relation of the photographed surfaces to the light sources in the scene and shadowing. This can be partially solved by using illumination compensation techniques [25, 26, 27].

2.2.3 Multi-view plus depth coding

An MVD video sequence can be encoded using methods for multi-view video coding where the color sequences and the depth sequences are encoded as separate multi-view sequences. The depth data differs statistically from the color 2D video sequence

due to its slow changing surfaces and discontinuities at object borders[28]. In [29] Fehn et al showed that the depth data can be compressed so that it contains only 10-20 % of the data in the color data, due to slow changing surfaces. However, it was also shown in [30] that the low quality of the discontinuities in the depth map has a high impact on the quality of synthesized views. Hence, the strength of the compression should be limited to avoid severe artifacts at the discontinuities when the standard H.264 and its extension MPEG-C part 3 are applied

It is possible to compress the depth maps more efficiently with sufficient quality if these statistical differences are considered in the encoding. One method is to include both depth and color data in the estimation of the motion vectors [31]. Then one set of motion vectors can be used to encode both depth and color data in a 2D plus depth sequence. In [32] Pourazad et al use the depth data characteristics to adapt the mode selection of the motion vectors in H.264. Approaches that encode edges in the depth data separately have been proposed in [33, 34, 35]. Krishnamurthy et al in [33] applies JPEG2000 based ROI image coding where depth discontinuities are detected as ROI. A scheme that segments and encodes the edges and main objects in images was proposed in [34]. A video based approach for segmenting and encoding of edges is found in [35].

Compression of MVD has been improved by using platelet-based depth coding [36]. The result is a higher rendering quality for this scheme than for H.264 intra coding of depth images. The interview redundancy of depth data has been addressed by Ekmekcioglu et al in [37]. They propose a temporal sub-sampling scheme to remove frames that can be successfully reconstructed using data from neighboring views. The previously mentioned approaches aim to adapt the encoding. A scheme that applies pre-processing in the form of adaptive smoothing of the depth data [38] has also been suggested.

2.3 Rendering

A representation that includes depth data and/or data of occluded objects can be chosen to reduce the number of views to transmit and therefore reduce the total bit rate. The views that are not transmitted can then be rendered from the provided data. Another advantage is that one transmitted 3D video sequence can be used for different types of 3D displays and is not limited to the camera positions of the capturing process. In this thesis the focus is on multi-view plus depth video sequences. Hence, only view synthesis based on multiple views and depth is addressed in this section.

Three problems have to be addressed in the synthesis process assuming that the

camera parameters of the available views are known and the camera position of the new virtual view has already been chosen. Firstly, the data from the available views must be transformed to the 2D plane of the view to be synthesized. This can be achieved using 3D warping. (See section 2.3.1.) In most cases several of the views of the multi-view plus depth sequence are used in the synthesis process to reduce disocclusion artifacts. Hence, the next step is to merge the views. (See section 2.3.2.) The last problem is to address holes in the final view due to round off errors in the 3D warping, disocclusion and other artifacts, for example due to variations in illumination between the views. (See section 2.3.3.)

2.3.1 3D warping

The depth map and the camera parameters of the available view are all that is required to project the color data from this view to a new virtual view. The camera parameters are divided into two groups; the extrinsic and the intrinsic parameters as described in [29, 39]. The extrinsic parameters describe the translation and rotation required to transform a 3D point $\mathbf{M} = (x, y, z, 1)^T$ of the scene into a pixel $\mathbf{m} = (u/z, v/z, 1)^T$ in the image plan of that camera, using homogenous coordinates. The extrinsic parameters are represented by the transformation vector $\mathbf{t}$ and the 3×3 rotational matrix $\mathbf{R}$. The intrinsic parameters describe the characteristics of the camera that influences how each pixel in the camera coordinate system is mapped to the image plane. They are represented by the calibration matrix $\mathbf{K}$ and contain information about focal distance, image centre coordinates and pixel sizes of the camera.

In order to create a virtual view from one or more available views, the relation that two points that portray the same 3D point $\mathbf{M}$ in the scene would have the same color value can be used. This idea is used in the process of warping one pixel $\mathbf{m}_L = (u_L/z_L, v_L/z_L, 1)^T$ from the left view to a pixel $\mathbf{m_v} = (u/z, v/z, 1)^T$ in the desired virtual view, as can be seen in figure 2.8, where the coordinate system of the left view in homogenous coordinates is $(u_L/z_L, v_L/z_L, 1)$. The same can be applied to generate a pixel based on the right view in figure 2.8. This process is called forward warping. The projections of a point $\mathbf{M}$ in 3D space to a point in the image plane of a camera $\mathbf{m}$ is defined by

$$z\mathbf{m} = \mathbf{PM} \tag{2.1}$$

where z corresponds to the depth value of the view described by the projection matrix $\mathbf{P}$ of that camera [39]. In the case as in [29] the projection matrix is further defined as $\mathbf{P} = \mathbf{K}\mathbf{I}_{3\times4}\mathbf{D}$ where the 4×4 matrix $\mathbf{D}$ contains the rotational matrix $\mathbf{R}$ and the

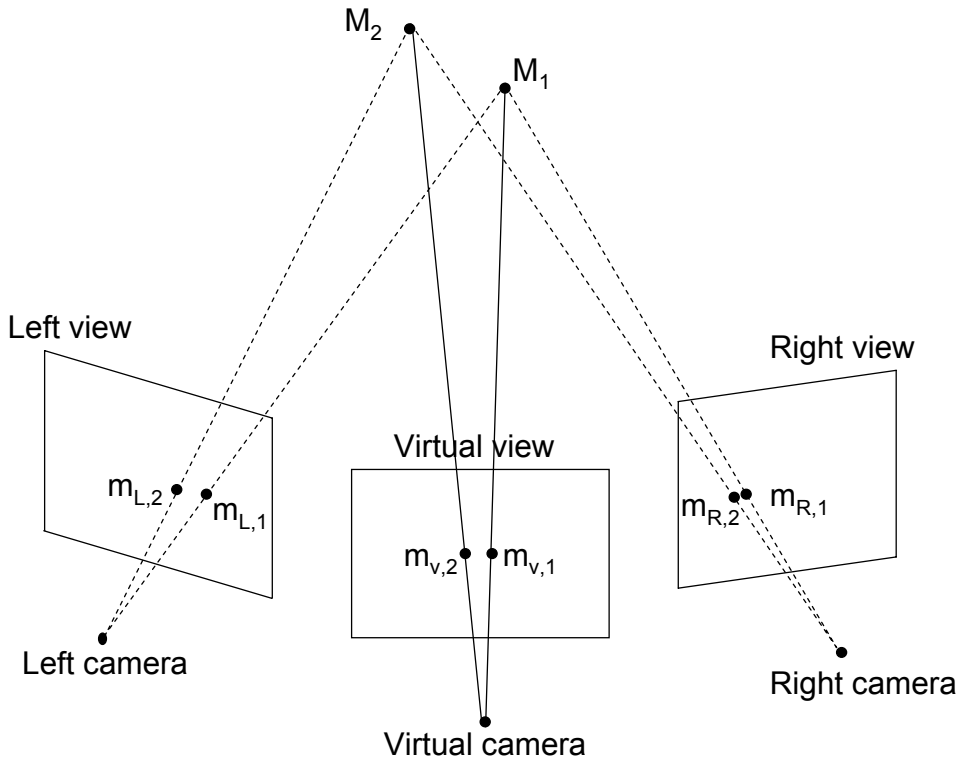


Figure 2.8: 3D warping of one pixel from either the left or the right view to the center view.

translation vector $\mathbf{t}$.

$$\mathbf{D} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Hence, the end points of the image warping are defined by

$$z_L \mathbf{m_L} = \mathbf{P}_L(\mathbf{M}) \tag{2.2}$$

$$z_v \mathbf{m_v} = \mathbf{P}_v(\mathbf{M}). \tag{2.3}$$

The point in space can then be expressed using eq. 2.2 as

$$\mathbf{M} = z_l \mathbf{P}_L^{-1}(\mathbf{m}_L) \tag{2.4}$$

Then the relation between the two points $\mathbf{m}_L$ and $\mathbf{m}_v$ is given by substituting equation 2.4 into equation 2.3 which gives

$$z_v \mathbf{m_v} = \mathbf{P}_v \mathbf{P}_L^{-1}(\mathbf{m}_L) \tag{2.5}$$

The answer to equation 2.5 rounded off to the nearest integer gives the coordinates (u, v) in the virtual view that corresponds to the coordinates (u_L, v_L) of the left view in figure 2.8. Hence, the color of the pixel at $(u/z, v/z, 1)$ is equal to the color of the pixel at $(u_L/z_L, v_L/z_L, 1)$. The exception is when more than one of the points from the original view is warped to the same point in the virtual view. This occurs when one of the corresponding 3D points occludes the other or when the space between two points is smaller than the size of a pixel. The visibility problem in the first case can be solved by calculating the depth values of the virtual scene or by applying the more effective occlusion-compatible warp order, which only depends on the position of the two cameras [29]. An alternative to only warping points is to warp point splats, where each new point in the virtual view is represented by a sphere. Thus, each point can influence a small part of its neighborhood depending on the distance to that point. The advantage is that small holes due to round off errors and small disocclusions do not appear; only larger disocclusions remain a problem. However, this approach increases the complexity of the warping and applies the same method when solving two types of errors that have different characteristics.

Another possibility would be to use the equations above to perform a backward warping instead, where each pixel in the virtual view would be determined by finding the corresponding pixel in the original view. However, this process would be rather complex, since it is becomes necessary to search the whole view to find all possible sources for that pixel [40].

2.3.2 Merging of the views

When multiple views and depth maps are available, the 3D warping in section 2.3.1 can be used to create candidate views. More than one candidate view can be used if several original views are available. Two candidates are in most cases created from the two original views at the least distance from the position of the desired virtual view. The next step of the rendering process is then to merge these candidate views into one single view. This process minimizes the rendering artifacts compared to using data from one only view. The main problem to solve involves what data to use regarding the case of the available information from both views.

The possible solution is to prioritize pixels that originate from the closer of the two original views. Another approach presented in [41] involves a sampling density measure (SDM), which is used to decide which view has the higher priority for each region. The SDM indicates how dense the warped pixels from a view are positioned in the virtual view, since a sparsely sampled virtual view has a reduced quality due to missing pixel values. Cooke et al in [41] also proposes to blend the color values $I_L(u_L, v_L)$ and $I_R(u_R, v_R)$ of the two views using a bilinear interpolation based on

the distance to virtual view. The color value of the pixel at (u, v) in the merged virtual view then becomes:

$$\mathbf{I}(u, v) = \alpha I_L(u_L, v_L) + (1 - \alpha) I_R(u_R, v_R)$$

where α is a linear weight. A similar approach has also been suggested in [42] where the blending is used for all pixels where a value in the warped image exists for both (u_L, v_L) and (u_R, v_R) and the corresponding depth values are close in value. The weight α depends in this case on the distance between the original view and the virtual view using the translation vector.

$$\alpha = \frac{|\mathbf{t} - \mathbf{t}_R|}{|\mathbf{t} - \mathbf{t}_L| + |\mathbf{t} - \mathbf{t}_R|}$$

The quality of the result from the bilinear filtering of the pixels from both original views depends on the correspondence of the two pixels. The bilinear interpolation may cause blurring if the two views are warped to a slightly different position in parts of the view.

2.3.3 View synthesis artifacts

The final virtual view that has been produced by first warping several views and then merging them together suffers from several types of rendering artifacts. The exception is when the disparity between the views is small. One problem is that neither the sampling density nor the position of the samples of the warped view corresponds to the pixel grid. Hence, cracks will appear in the virtual view as can be seen in figure 2.9. In [41, 42] the virtual view is median filtered, which does not remove details as the blurring filter. The disadvantage of using a median filter is that the edges in the image can become unnaturally sharp. A solution is to only apply the median filter to the cracks as in [40].

The next problem involves holes in the image where data is missing due to disocclusions. (See figure 2.10.) This problem can either be solved by estimating the missing data from the available data or by preprocessing the original color and depth data to reduce the effects of the disocclusion. The missing data can be estimated using a simple approach such as bilinear interpolation [41] or to extrapolate constant color background pixels [40]. The background pixels are those that are further away from the camera. Another common method is inpainting [43], which uses the information of the surrounding pixels and their characteristics to provide an improvement to interpolation-approaches. The inpainting can be further improved by using data from multiple views [44] or to use the depth data only to include information in the background [45].

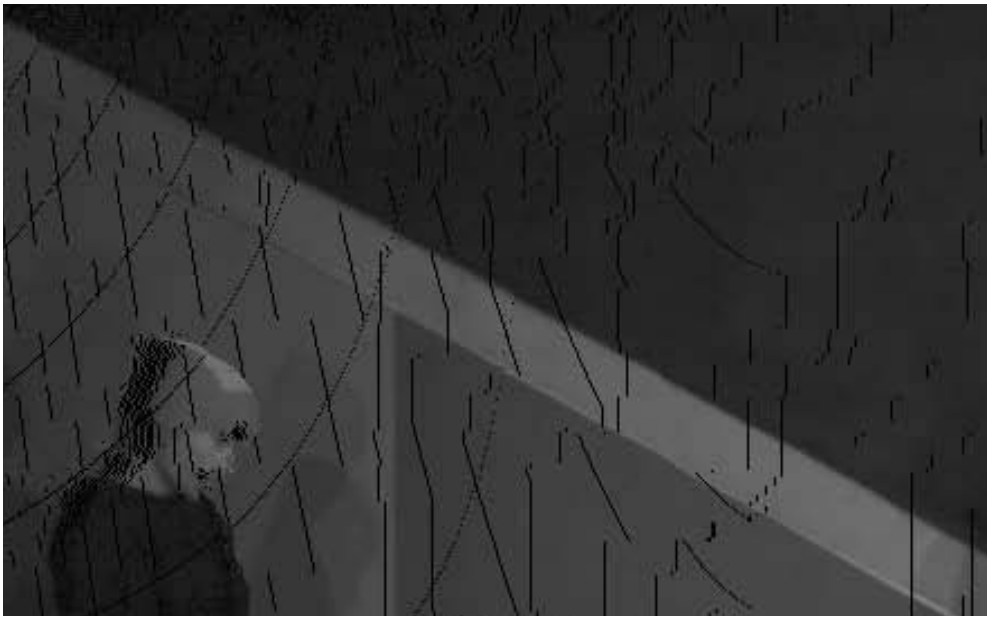


Figure 2.9: Cracks in the warped image occur since the sampling density or the position of the pixels does not match the pixel grid.

The disocclusions occur close to object boundaries, which are represented in the depth map as depth discontinuities. The smoothing of the edges with for example Gaussian filters as proposed by Fehn in [29] removes small disocclusions at the expense of artifacts in the color map due to the distorted depth data. A solution is to restrict the smoothing only to the edge region [46, 47] and in [48] only in the vertical direction in order to minimize the distortion. Park et al in [38] take this a step further and propose an adaptive filter that consists of both a discontinuity-preserving smoothing and a gradient direction-based smoothing. Another solution for removing small disocclusions and cracks is to use the point splat method during the image warping as mentioned in section 2.3.1.

The final problem mentioned in this section relates to the boundaries of two regions containing warped data that does not originate from the same view. Depending on the disparity, the difference in illumination and errors in the warping, the so called corona effects may occur near objects where a part of the background next to the boundary is taken from another view as can be seen in figure 2.11. Solutions include low pass filtering these unnatural edges [40] or to use color correction before warping to reduce the effects from illumination [46].

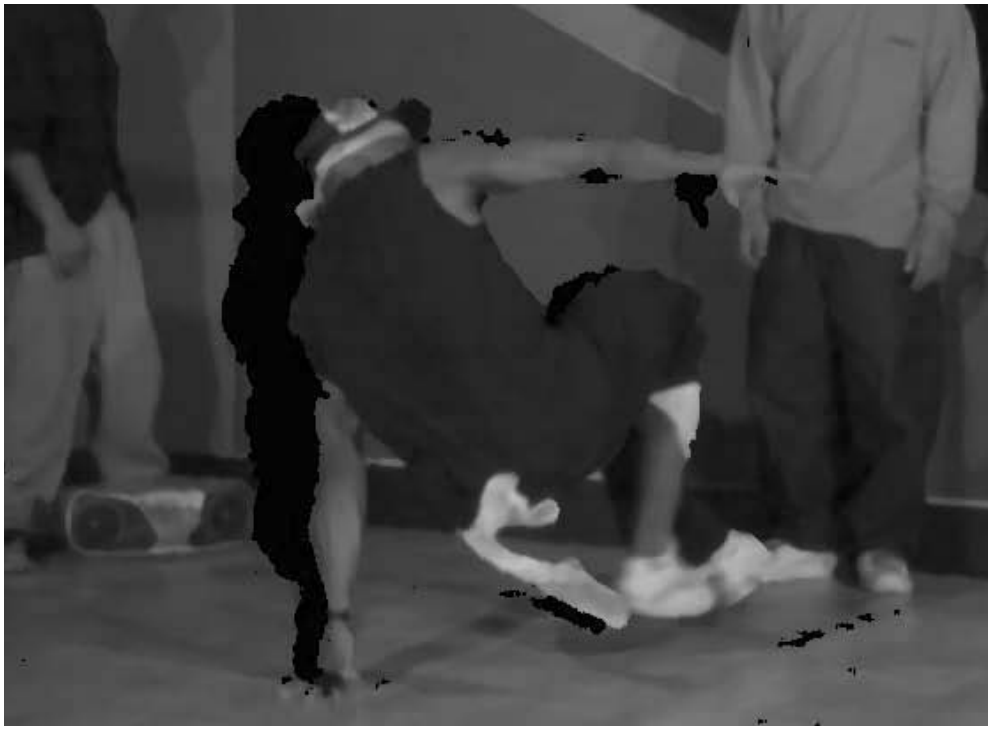


Figure 2.10: Large holes in a warped view are the result of dissocclusions.

2.3.4 Layered approaches to rendering

One of the main problems in the rendering is object boundaries where disocclusion might appear along other artifacts due to the depth discontinuities in the depth map. A solution is to partition the image in to a boundary layer and a layer containing the rest of the data as in [40, 35]. The regions are warped separately for both the left and the right original view and are thereafter merged into one view. In addition a three layer approach has been suggested by Huang et al in [49], where the data is divided into a main layer, a background layer and a boundary layer.

2.4 Video quality

Quality assessment of video is usually performed with fast and repeatable objective measures, which use predefined algorithms. However, the algorithms fail to include all aspects of the human visual system (HVS) and therefore do not give a precise measure of perceived quality. Subjective tests, using human subjects, give a good representation of the HVS. However, they are time-consuming and have high

Figure 2.11: The artifacts in the form of outline of objects in the background are called corona effects.

requirements with regards to the test setup in order to ensure analyzable tests.

Assessing the quality of 3D video is more complex than for 2D video due to its multidimensional nature. Another aspect is that the whole chain from capture/creation of 3D video to the actual display can affect the final quality as perceived by the viewer. There are a vast number of choices that can be made for each step making the quality assessment of isolated steps a difficult task.

2.4.1 Perception of depth

The key to a realistic 3D experience lies in the perception of depth by the viewer. The information necessary for depth perception is described by a set of depth cues. These include psychological depth cues where the HSV uses information in an image such as occlusion, lighting, shading and linear perspective to determine the depth of the scene. These can also be included in 2D video to provide a more realistic experience. There are additional depth cues that correspond to the physiological relation

between the depicted scene and the HVS. The binocular depth cue is used to determine the depth based on the disparity between the views visible to the right eye and the left eye. These two views are then blended by the HVS to construct one 3D image, in which the slight difference between the images provides the sensation of depth. Another depth cue, motion parallax, relates to the movement of the head with respect to the scene. If the head is moving, the position of the eyes changes in relation to the scene, which implies that two new images should be presented to the eyes to give the correct experience. Additional physiological depth cues include, accommodation, where the lens of the eye is shaped to focus on an object in the scene. The eyes are rotated so that the gaze is fixated on this object, namely convergence. If a 3D system does not comply with the depth cues, a loss of the 3D sensation and fatigue due to eye strain might occur. An additional effect that is typical for stereoscopic video and images is binocular suppression. In other words, only one of the two views requires high quality in order to perceive high quality.

2.4.2 Objective quality measures

Video coding quality is, in the majority of cases, measured using peak-signal-to-noise ratio (PSNR) [50], which is based on the mean square error (MSE).

$$PSNR_{dB} = 10 \log_{10} \frac{255^2}{MSE}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2,$$

where N is the number of pixels within the video signal, 255 is the maximum pixel value and x_i and y_i represents pixel i in the original and distorted video sequence, respectively. This measure only considers errors at the pixel level completely disregarding the position of the pixel or the context of the region. ROI video coding aims to improve perceived quality by increasing the quality in the ROI at the expense of that in the background, and therefore the PSNR of the ROI is often analyzed separately. This gives an indication of how much the quality increases within the ROI, but no indication of the effect of the decreased background quality to the overall perceived quality. Attempts have been made to include more HVS characteristics in objective measures, which are summarized by Wang et al. in [51]. Several approaches have been considered such as applying foveation and visual attention [52], contrast sensitivity [53], structural similarity metric (SSIM) [54], video quality metric (VQM) [55] and others.

Quality in 3D video is still largely evaluated using 2D quality metrics despite the fact that those metrics do not include any information about depth perception,

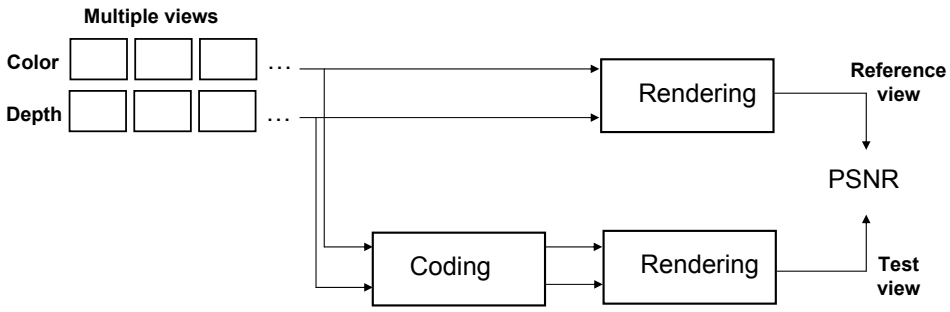


Figure 2.12: PSNR metric of virtual views of a multi-view plus data sequence. The comparison of rendered virtual views of both the original and encoded data reduces the impact of rendering artifacts.

stereoscopic distortions, immersiveness, naturalness or comfort for example. In most of the papers addressing compression of multi-view plus depth video, PSNR has been used to determine 3D video quality. PSNR can be applied directly, as for the 2D case, if all views required by the application are compressed and available. However, in the case of multi-view plus depth video compression the rendering step is included in the quality assessment. The compression of depth data and to some extent the color data may introduce rendering artifacts that should be taken into consideration. In this case, an average PSNR is calculated by comparing views that are synthesized from compressed data to views that are synthesized from the original data as can be seen in figure 2.12. By doing so, the effect of the rendering is strongly reduced.

In [56], Olsson et al. proposed to use the PSNR value of the color data at certain depth intervals of 3D images based on integral imaging. This takes the relation between the position of objects in the scene and the perceived quality into consideration. Other metrics that include effects of depth perception include the PSNR metric for stereoscopic video that weights the left and right view based on binocular suppression [57]. Binocular suppression is the ability of the HSV to disregard errors in one of the views, if the other view has high quality. In addition, in [58] Benoit et al. include disparity between two stereoscopic views to measure the effect on depth perception.

2.4.3 Subjective quality measures

Subjective tests on humans can be used to measure the actual perceived quality. Successful extraction of analyzable results from subjective tests depends on several fac-

Vote	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2.1: The ITU-R Quality Scale used to evaluate the subjective quality of one video sequence.

tors including the test time, test order and the instructions received. In addition, even if one test person experiences good quality, this is not necessarily the case for the next test person. The standard ITU-R BT.500-10 [59] presents several different test methodologies concerning quality assessment of single video sequences and pair wise comparison of 2D video sequences among others. The definitions include setup parameters and a description of tests and scales that can be used for a quantitative evaluation of the perceived quality. The ITU-R quality scale in table 2.1 is used for evaluating single sequences, while the scale in table 2.2 can be applied to the comparison of two video sequences. Subjective tests with a few subjects and/or few test sequences can be used to verify the results of objective measures.

The complexity of the subjective test increases further when 3D video is considered. The evaluation of 3D video depends highly on the technology used to present the data to the viewer. The large variety of display techniques and the various artifacts that the screens generate increases the problem of interpreting the results of the tests. In particular, if a compression or rendering technique is evaluated that is not specific for one display. Subjective tests of 3D video can also be evaluated using parts of the standard ITU-R BT.500-10 [59] and in particular the quality scales. This can be combined with the recommendation ITU-R BT.1438[60] that for example defines some additional assessment factors, viewing and how to check if the test subject has the ability to perceive binocular disparity. Currently there is no standard or recommendation that directly addresses subjective assessment of multi-view video.

Applying the standard enables comparison with other research results.

Vote	Quality
-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
1	Slightly better
2	Better
3	Much better

Table 2.2: The ITU-R Comparison scale used to evaluate the quality of one video sequence compared to another.

Chapter 3

Region-of-interest coding

One of the problems addressed in this thesis involves how to transmit video at low bit rates with as small an impairment as possible on the perceived quality. Region-of-interest (ROI) video coding can be applied to achieve a higher quality within the ROI at the expense of the quality in the background, if the correct ROI can be detected. The ROI coding is performed in two separate steps. Firstly the ROI is detected by predicting the type of content in a region that attracts the viewer's gaze and communicates the greatest amount of information. Based on these characteristics the position of the ROI is extracted (See section 3.1). In the second step, the bit allocation is controlled in order to ensure that more bits are allocated to the ROI to increase the visual experience (See section 3.2).

ROI coding can be used with a user-defined ROI or automatically detected if the important parts can be successfully predicted. For example, the information communicated by the movement of the lips and facial expressions in a video conference is lost if the artifacts are sufficiently large. In addition, reduced quality in the facial area can appear more disturbing for the viewer than reduced quality in the background. It was shown in [61] that high quality of the face is particularly important for people with hearing impairments. Several approaches were shown to improve the quality of video containing talking humans by applying more compression to the background than in the facial region, including approaches by Eleftheriadis et al. [62] and Chen et al. [63].

In surveillance applications the regions of interest are often well defined, for example people or vehicles. Examples include detecting camouflaged enemies in the shape of people or military vehicles [64], traffic surveillance [65] and security applications [66].

In [67], McCarthy et al. observed that the quality of the player and ball in a soccer

game proved to be the most important factor for the viewer. Detection of ROIs in soccer sequences has also been addressed by Kang et al. in [68].

3.1 ROI detection

The key to a successful ROI video coding is to correctly predict and detect the ROI, since a falsely detected ROI gives a lower perceptual quality than for ordinary video coding. The ROI is detected using either a generalized or an application-based approach. The generalized approach is based on visual attention models presented in section 3.1.1. In the application-based approaches the type of content present in an interesting region is predicted a priori for a particular application. These include video conferencing and videophone applications (where faces are of interest), surveillance of people and vehicles as well as sports applications.

3.1.1 Visual attention

A visual attention model determines the likelihood that a human viewer fixes his/her gaze on a particular position within a video sequence. This is generally based on models of the HVS. In [69] and [70] metrics of color, orientation, direction of movement and disparity are combined into a saliency map indicating the probability of this pixel drawing attention. Bollman et al. in [71] and Ho et al. in [72] extended these approaches by including motion detection. In additon, Ho et al. in [72] also includes face detection. These approaches assume that the visual attention models can be generally applied. However, Ninassi et al. in [73] showed that the positions at which people are gazing are affected by the given task. This implies that the detection would be more accurate when the task is considered in detection, which is often the case in application-based approaches.

Face detection

In video communications, human faces communicate a substantial part of the unspoken information using facial expressions and lip movements. The interpretation of the information becomes more reliable when the quality of the face is improved by the ROI coding. Another effect is an experience of increased quality of the viewer. Face detection is a popular area. It includes research on face recognition for surveillance and biometric authentication, coding using 3D-models, and applications, where facial expressions are extracted and analyzed. In [74], Yang et al. present an overview of methods used for face detection including feature-based

methods, template matching, eigenfaces, neural networks, support vector machines, and more.

Skin-color detection methods are often used in ROI video coding, where the face is considered interesting to the viewer. Skin-color detection is a fast method, which is invariant to pose, orientation, difference in skin-color and occlusion. An overview of pixel based skin detection techniques can be found in [75]. A wide variety of color spaces have been applied including RGB, normalized RGB,TSL, YCbCr and others, where different color maps are suitable for different purposes. YCbCr is a common choice, since it is used in most video compression standards and separates luminance from chrominance. YCbCr can be calculated from RGB using the simple transformation

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ C_r &= R - Y \\ C_b &= B - Y \end{aligned} \tag{3.1}$$

Several methods can be used to classify skin color based on a particular color space. These include the parametrical models such as skin cluster boundaries [63, 76, 77], elliptic models [78], single Gaussians models [79, 80] and Gaussian mixture models [81].The non-parametric approaches such as Bayes classifiers [82, 83] gives a better performance at the cost of increased storage space and higher requirements on the training set.

The detection of skin-color can be combined with other feature detection methods to ensure robustness with respect to illumination changes and increased selectivity. These features include eyes and face boundary in [84] and [77], mouth in [84], size of ROI and number of holes in the region in [80], and a homogeneity metric in [83, 85].

Other face detection techniques for ROI video coding include extracting facial contours from edges in [86] and using a unsupervised neural network described as a self-organizing map (SOM) as in [87].

3.2 Bit allocation

Reallocation of bits from the background to the ROI can be achieved either by a pre-processing stage before the encoding, or by controlling the parameters in the encoder. The video coding standards allow alterations of the encoder as long as the required features are included and the syntax of the bit stream is unaltered. Modifications of all encoders when the ROI algorithm is changed can be avoided by using

pre-processing. The adaptability of the background quality to variable bit rates of the channel is then reduced.

3.2.1 Spatial bit allocation

The majority of the research on bit allocation for ROI video coding applies spatial methods and in particular those controlling parameters within the codec. The quantization step sizes control the quantization accuracy concerning DCT components and the prediction error. In [88] and [63] two step sizes are used for each frame. A small step size is used for the ROI and a larger step size is used for the background. This results in a more finely tuned quantization and therefore improved quality of the ROI. However, the difference between the background and ROI quality appears abrupt if there is a large difference in quantization.

Solutions include adaptation of the quantization step sizes of the background to the distance of the ROI border [89, 90], applying three quantization levels in [86] or naking decision relating to the step size based on a sensitivity function [91]. Approaches involving other types of encoder parameters are addressed in [92] and [93], where the number of non-zero DCT components is used to control bit allocation. YLiu et al. in [94] controls the bit allocation by adapting several types of encoder parameters. These include quantization parameters, mode decisions, number of reference frames and search area for motion vectors. Sivanantharasa et al proposes in [95] the use of a tool for error resilience coding in H.264, namely flexible macro-block ordering, for ROI video coding.

Controlling quantization parameters allows direct integration with the rate-distortion function within the codec, but can introduce "blockiness" due to coarse quantization. Pre-processing, whose resulting error is generally less disturbing, avoids codec dependencies. Methods based on low-pass filtering (blurring) of non-ROI are addressed in [63] and [96]. Low pass filtering reduces the amount of information which gives less non-zero DCT components and a reduction in the prediction error due to the absence of high frequencies. The rate-distortion optimization of the encoder then re-allocates bits to the ROI, which still contains high frequencies. In [63] Chen et al. applies low pass filtering of the background regions, where one filter is used for the complete frame. This gives a distinct boundary separating the ROI and background which leads to disturbing artifacts. The foveation coding approach in [96] uses a fixation point at which the human is predicted to gaze. A gradual transition of quality from the fixation point to the background is then achieved using a Gaussian Pyramid. Foveated Coding is addressed in section 3.4.

3.2.2 Temporal bit allocation

The number of bits necessary to encode the background is reduced if a lower frame rate (fps) is used, which causes a decrease in the background quality. The related object-based video coding addressed in section 3.5 extracts and encodes objects and background in separate layers. These layers are synthesized into one video sequence at the decoder. Reduced frame rates for the background as in [97]can be achieved by encoding the layers using different frame rates. Object-based coding is supported by the MPEG-4 standard [7]. A similar approach is suggested in [65] by Meessen et al where the ROI and background are encoded and transmitted as two separate sequences. Such a technique requires adaptation at the receiver side.

Compatibility with the standard is preserved when the syntax of the bit stream remains unaltered. Temporal bit allocation schemes that are compatible with standards are proposed in [86, 98, 99]. In [98], all blocks not used in the encoding of the ROI are skipped in the P-frames. The DCT coefficients of these blocks are deleted in the I-frames. Lee et al. in [86] reduces the transmitted information by skipping background macro blocks in every second frame unless the global motion in the frame exceeds a threshold. A similar approach is presented by Wang et al. in [99], where the background blocks are skipped based on the content in the ROI and background.

Adiono et al. presents a pre-processing approach in [92] that applies a temporal average filter in order to average out the differences between the background of two frames.

3.2.3 Combinations of spatial and temporal bit allocation

The spatial methods reduce the background information transmitted in DCT components or the motion prediction error. The temporal filters, on the other hand, reduce the bits assigned to the background motion vector. (The exception is the average filter in [92], which affects the prediction error instead.) Hence, combinations of spatial and temporal approaches increase the reallocation of bits from the background to the ROI. Lee et al. in [86] combine the control of quantization step sizes, with the skipping of background blocks for every second frame. The background blocks are only skipped during limited global frame motion. A similar approach as in [86] is proposed by [99], with the difference being that the skipping of background blocks is adapted to ROI and background content. Adonio et al. uses a temporal average filter in combination with controlling the quantization step sizes.

3.3 3D ROI video coding

ROI video coding can be extended to 3D video, in which the depth is a factor considered in the detection of the ROI [100, 101]. It is assumed that objects close to the viewer are of importance. The knowledge of the geometry available in synthetic scenes can be used to calculate the ROI. In [100] Cheung et al proposes a scheme that detects the distance to the camera and the position of the edges of objects. This is achieved using values in the Z-buffer during the mapping of 3D objects on to the 2D plane. The extracted information is used in order to weights in the assignment of bits to each MB in H.263+. Masala et al in [101] apply a similar scheme using an MPEG-2 codec in a system that keeps track of user interaction and thus the position of the viewer.

ROI image based coding has been applied to stereo images. A region-based scheme was proposed by Ayudinoglu in [102] using three types of regions; occlusions, edges and smooth regions. Margues et al in [103] showed that the visual attention methods for 2D ROI image coding are improved by incorporating depth information in ROI stereo image coding. ROI coding has also been applied to improve the JPG2000 compression of the depth map of a 2D plus depth video/image in [33]. The choice of ROI is based on the rendering process. The compression of the edges may cause artifacts at the discontinuities. Hence, the edges are included within the ROI.

3.4 Foveated coding

A related research area to ROI video coding is to use fovea instead of ROIs. In biology, the fovea is the part of the retina in the human eye that contains the greatest number of photoreceptors. Details in images can only be perceived if that part of the image is processed by the fovea. Thus, only the point upon which the human gaze is currently fixed must be presented with good quality, which enables quality reduction based on the distance to this point. The foveas are placed with their centers at the pixels where the human is predicted to gaze in each frame as in [96]. This approach demands that the exact location of a person's gaze is known, whereas in ROI coding, it is only necessary to detect the region of the gaze.

3.5 Object-based coding

In the standard MPEG-4 [7], objects and background can be divided into a set of layers. The layers are compressed separately and then synthesized into one sequences

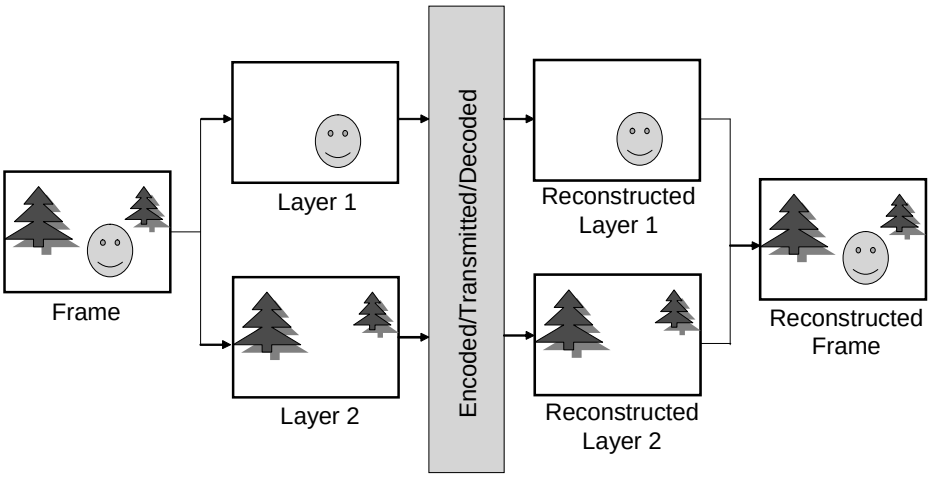


Figure 3.1: Object-based coding. The different objects in a frame and the background are encoded in separate layers. The reconstructed layers are synthesized into one frame after the decoding of the layers.

at the receiver. The partition into background and object layers is described in figure 3.1. ROI coding differs from object-based coding in that regions of different characteristics are extracted instead of specific objects. In addition, most ROI coding approaches do not use layers. The sequence is transmitted as an ordinary video sequence.

3.6 Contributions

The author has published four papers in the field of ROI video coding, where papers I-III address pre-processing of 2D video in the form of filters in both the spatial and temporal domain. A codec independent spatio-temporal filter is proposed in paper II, which combines the spatial filter in paper I with a temporal filter. A theoretical analysis and quantitative tests using objective quality metrics in papers II and III was used to analyse the coding efficiency, the reallocation of bits from the background to the ROI and the complexity of the filter. The effect of the choice of filter parameters on rate-distortion was analysed in paper III. A minor subjective test was applied to verify the analysis of the SPTP filter. The spatio-temporal filter is then extended in paper IV to provide ROI coding of 3D video in the 2D plus depth format using the depth data in the detection of the ROI. Each of the four papers is described in more detail in the rest of this section.

3.6.1 Paper I

A spatial (SP) filter is proposed in paper I that improves the ROI approach in [63] by allowing a gradual degradation in quality from ROI to the background. The gradual transition in quality reduces the border effects. These appear when the SP filter causes large differences between neighboring ROIs and background pixels at the ROI borders. This is achieved by extending the idea of using different Gaussian filters depending on the distance to the border of a region instead of the distance to a point of fixation, as in the foveated coding approach presented by Itti in [96]. ROI video coding does not detect the exact position of the human gaze and therefore equal quality within the ROI is necessary. The position of the ROI is detected using a simple skin-color detection algorithm; the parametric model in [78] with experimentally determined thresholds.

The next step is to match each pixel in the background to the appropriate Gaussian filter. This is achieved by using quality maps, which are created by low pass filtering the result of the ROI detection. The position of the ROI and the distance from the border are then extracted from this map in the SP filtering. Tests using PSNR were performed on the carphone sequence. The gradual quality reduction in the background was shown not to reduce the quality in the ROI compared to ROI with only one filter as in [63].

The issue of computational complexity of the SP filter was also addressed. An alternative approach that combines the ROI detection with a variance metric was proposed and tested. The tests showed that more than a quarter of the background pixels could be omitted in the filtering without a noticeable quality reduction within the ROI for the tested sequence. This result can be extended to most video sequences, since areas with only low frequency components are in general found in frames of a natural video sequence.

3.6.2 Paper II

ROI video coding approaches that operate within the codec to reallocate bits from both the spatial and temporal domain to the ROI are proposed by [86, 99]. In addition, Adiono et al in [92] proposes to combine a pre-processing in the form of a temporal average filter with a method that controls the number of DCT components within the codec. The utilization of both the temporal and spatial domains can increase the coding efficiency substantially. In paper II a codec-independent spatio-temporal solution is proposed. The proposed filter is a combination of the SP filter of paper I and a TP filter. The TP filter extends the codec-dependent approach in [86], which can transmit the background at a lower frame-rate than the ROI. A sim-

ilar result is achieved by the TP filter that removes all changes in the background in every second frame. Bilinear interpolation of the part of the background close to the ROI border is used to reduce border artifacts. More details on the temporal filter can be found in [104].

The influence of the TP filter on motion vector assignment and prediction error was evaluated. The theoretical evaluation showed that the cost in bits of motion vectors decrease in the background whereas the bits assigned to the prediction error remains the same. This indicates that fewer bits are required to encode the background. However, it also implies that the non-required bits are assigned to improve blocks with large prediction errors in both the background and the ROI.

The qualitative analysis of the assignment of bits to the background and the allocated of any additional bits by the encoder was performed for a SP or TP filtered video sequence. The evaluation showed that the combination of the two filters increases the coding efficiency compared to using either the SP or the TP filter separately. Any additional bits are assigned to improve blocks with large prediction error, which exists in both the background and the ROI for the TP filter. The combination with the SP filter solves this reallocation problem. An analysis of the computational complexity showed that the spatio-temporal (SPTP) filter has a lower complexity than the SP filter from paper I. This is valid unless the transition region occupies more than 75 % of the background.

The results in the qualitative analysis are verified by tests on sequences that were SP, TP or SPTP filtered. The filtered sequences were encoded using either both an H.264 or MPEG-2 encoder with fixed encoder parameters or a target bit rate. The test also showed that the difference in coding efficiency of the SP and SPTP filter depends on the motion content of the ROI. The SP filter was shown to provide a minor improvement compared to the SPTP filter in those sequences where the ROI has high motion content relative to the background. An explanation is that the context-adaptive binary arithmetic coding (CABAC) in H.264 adapts the motion vector lengths to favor the original data in these cases.

3.6.3 Paper III

The description of the SPTP-filter in paper II and the qualitative analysis regarding how the filtered data influences the encoding are extended into more detail in paper III. The computational complexity of the SPTP filter is compared to the alternative SP approach in paper I. The SP filter applies variance in the creation of the quality map to decrease complexity. The SPTP filter is shown to have a lower computational complexity than the SP filter if the transition region of the TP part of the filter does not exceed the total number of filtered pixels. The impact of the variance of the SPTP

filter was evaluated both theoretically and by rate distortion tests on the H.264 and MPEG-2 codecs. The evaluation showed that the variance controls the number of DCT components in the encoding. Hence, the bit rate can be adapted to the conditions of the channel, provided that the information is available.

A quantitative motion vector analysis confirmed that the SPTP filter does reduce the number of motion vectors assigned by an H.264 encoder. The lengths of the motion vectors are only slightly increased. The analysis indicated that the SPTP filtered sequence has a slightly lower coding efficiency than the SP filtered sequence when the ROI contains the main part of the motion vectors. This verified some conclusions from paper II.

The PSNR quality measurement is widely used but it does not consider the type of information within a region or in what context the video sequence is watched. Consequently, a minor subjective test was carried out to give an indication of the impact of the SPTP-filters on perceived quality. The stimulus comparison method [59] with 10 test subjects was used. The result of the subjective tests implies that the SPTP filter does in most cases improve the perceived quality of the tested video sequences. However, it was also shown that the viewer may be distracted by the large reduction in background quality caused by a large variance of the SPTP filter.

3.6.4 Paper V

ROI video coding has previously been extended to 3D video in [100, 101] using the position of the viewer and scene information in a synthetic scene to control bit assignment within the encoder. The encoded material is then rendered from the scene. Other approaches include the addition of disparity estimation to improve the results of ROI detection using visual attention [103] for stereo images. The idea of using depth or disparity information in ROI detection is extended to ROI coding of 2D plus depth video sequences in this paper. Two ROI detection methods are proposed. In the first approach the statistical properties of the provided depth data are analyzed. The aim is to detect the depth values at which the front-most pixels are located. The second approach is a combination of detecting the front most pixels and the skin detection used in papers I-III. The result is used to create a quality map as in the previous papers. Thereafter, the SPTP-filter in paper II is applied to the color part of the 2D plus depth video sequence.

The two ROI detection methods increase the quality as measured by PSNR within the selected ROIs. A visual comparison showed that the combination of the face and depth detection offers the better quality of the two approaches. This implies that objects close to the scene are important, but that additionally, the context may increase the importance of a background objects.

Chapter 4

Multi-view plus depth scalable video coding

Consider a heterogeneous network where the end-users are equipped with 2D or 3D screens of various types. The problem addressed in this section involves how to provide a high perceived quality for all end-users in the network who are viewing a 3D video sequence. The heterogeneous network would require the transmission of a multitude of different versions of the same video sequence to provide the best quality under the limitations of each separate end-user. A solution is to use scalable video coding (SVC). SVC enables a flexible extraction of data from a video sequence and can therefore be adapted to the available resources of an end-user. SVC can therefore increase the overall quality in a heterogeneous network.

4.1 Scalability in 2D

Scalability can be applied in various ways and combinations. The three main types of scalability are temporal, spatial and quality (SNR) scalability, which are all described in the in the SVC extension of h.264 [11]. In *temporal scalability* (See figure 4.1), the frames of a 2D video sequence are encoded to enable flexibility in the choice of frame rate. In other words, it allows for subsets of the frames within a GOP to be extracted while skipping the rest. After the encoding, the frames are reordered to make it possible for the lower layers to be extracted first within each GOP. The prediction structure of the GOP must be chosen such that a frame is not predicted from a higher enhancement layer. Otherwise prediction errors may occur due to prediction from non-existing frames. One additional advantage of hierarchical b-frames is that their structure ensures that all prediction is from a frame of the same layer or lower.

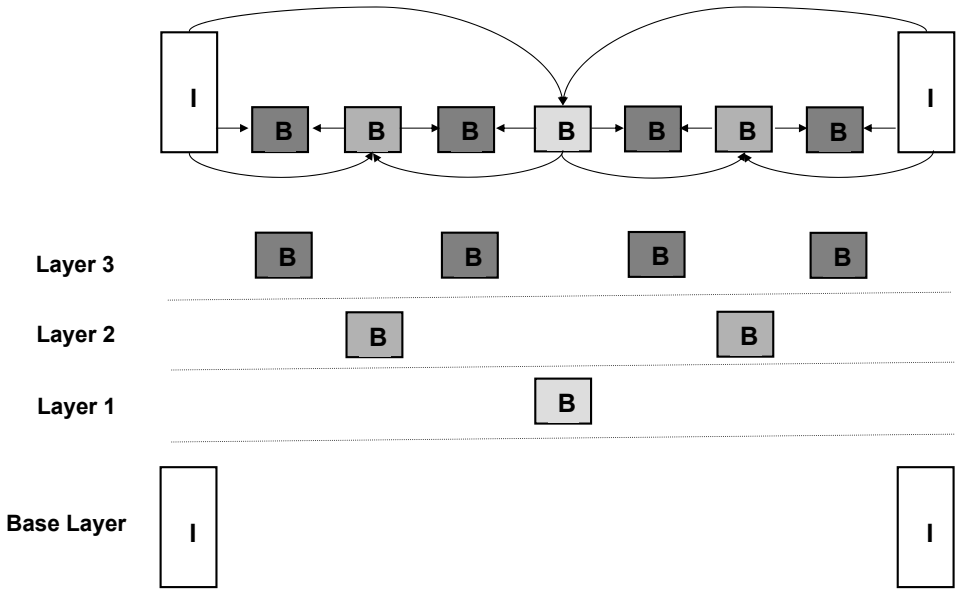


Figure 4.1: Temporal scalability. Each of the frames in the video sequence is assigned an enhancement layers such that the addition of one layer increases the frame rate.

Spatial scalability enables the extraction of a 2D sequence of variable spatial resolution from the SVC encoded sequence. The base layer contains a low resolution video sequence. Each of the enhancement layers contains a higher resolution version of the video in the previous layer. Interlayer redundancy is exploited by predicting each frame from an up-sampled version of the corresponding frame in the previous layer. Thus, the enhancement layers contain the data of that resolution not available in the previous layers. An example of spatial scalability can be seen in figure 4.2.

Quality scalability (or SNR scalability) addresses the quality within each frame. This can be achieved by allowing the quality of the sequences to be varied either in clearly defined layers CGS (coarse grain quality scalability), continuously as in FGS (fine grain quality scalability) or a mixture of the two MGS (medium grain quality scalability). (See figure 4.3.) In the first case, CGS, the data in each layer is treated in a similar manner to that for spatial scalability. The base layer contains a full 2D sequence at low quality and each of the enhancement layers contains high frequency components not present in the previous layers. FGS, on the other hand, has one base layer and one large enhancement layer. The data in the enhancement layer is encoded and arranged in such a way that the amount of enhancement data can be chosen continuously. This ensures flexibility in the adaption to the bit rate limitations of the transmission at the cost of either coding efficiency or drift errors. Predic-

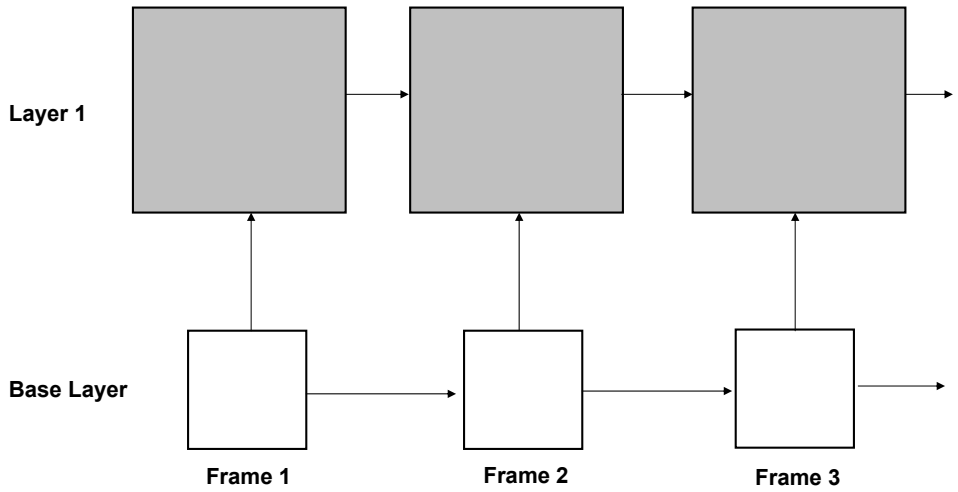


Figure 4.2: Spatial scalability. A spatially scalable video sequence consists of layers of different resolutions. In the encoding, each layer is predicted from an up-sampled version of the previous layer.

tive coding can be successfully used in CGS including the current and lower layers. However, in FGS it is not previously known how much of the enhancement layer will be extracted. Hence, drift errors might occur if prediction is used based on the enhancement layer. A compromise in the form of MGS provides an improved coding efficiency without drift errors compared to FGS and more flexibility in the choice of bit rate than CGS.

The use of the wavelet transform instead of the DCT automatically decomposes each frame such that spatial and quality scalability can be applied directly. The wavelet transform is not included in the SVC extension of SVC due to its increase in encoder complexity and reduced coding efficiency for video [11, 105].

4.2 Scalability in 3D

The scalability methods available in the SVC extension of h.264/AVC have been applied to multi-view video in [106]. The coding structure in hierarchical b-frames used in this paper facilitates the decomposition into layers of different frame rates (temporal scalability). The 2D scalability methods can also be applied directly to the multi-view part of multi-view plus depth and in [107] Cho et al suggested to apply spatial scalability to the depth data.

The multi-view and multi-view plus depth videos contain information captured

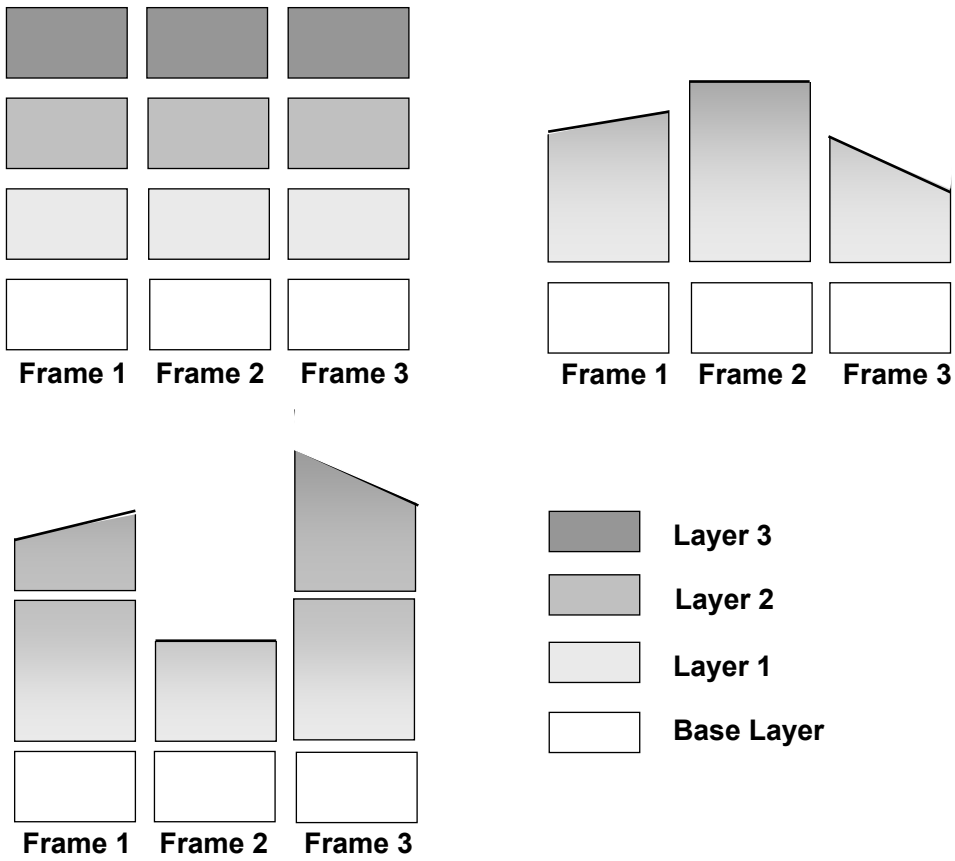


Figure 4.3: Quality scalability. The quality of the video sequence increases by each added layer. The layers can either be a) clearly defined, CGS, b) continuous, FGS or c) a mixture of the two, MGS.

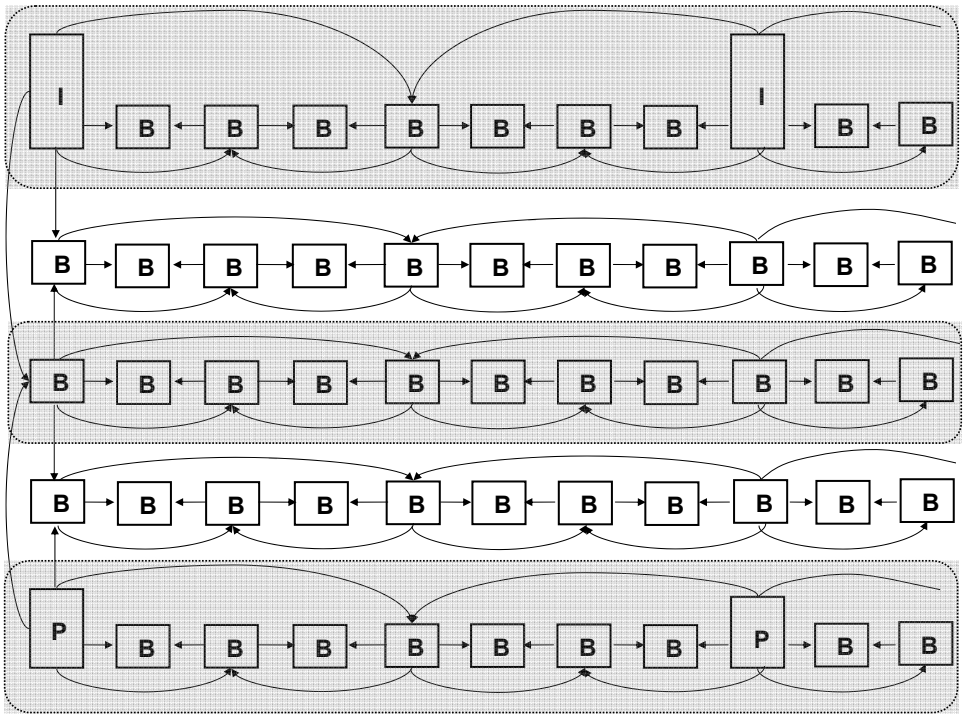


Figure 4.4: View scalability is enabled by using predictive encoding structures that allow extraction of a subset of the views. For example every second view could be extracted in this case where hierarchical b-frames are applied.

from several camera positions and by using various types of 3D techniques, this data can provide an experience of depth. The 3D characteristics of the data can also be utilized to provide additional scalability schemes. Scalability in the view domain can be provided using a similar approach to temporal scalability. Instead of extracting a subset of frames [22] Lim et al. propose extracting a subset of the views. (See figure 4.4.) The approach by Shimizu et al [108] provides a scalable solution that uses color data, depth data and residuals. The approach includes one view in the base layer, including the depth data corresponding to that view. The depth data necessary to transform this view into a view at another camera position are found in the enhancement layers. In addition, the residuals of the transformed views and the original views are included in the enhancement layers.

The depth ranges of multiscopic 3D displays are limited. Hence, artifacts appear when detailed information is displayed outside of the depth range of a display. The high frequency components causing the artifacts can be removed using filters. In [109] Ramachandra et al, suggest a depth scalability scheme that can adapt to the

depth range of a display.

Wavelet transforms have also been suggested to enable temporal, spatial and quality scalability [110] and in some cases even view scalability [111, 94]. However, the wavelet transform still has some problems that reduces the coding efficiency compared to block-based approaches [23].

4.3 Contributions

The author has published three papers (Paper V-VII) in the field of multi-view plus depth scalable video coding. An approach that provides a depth-based scalability of the color data is proposed in paper V. The depth scalability is combined with view scalability. Three different approaches to extract the layers from a view are proposed and analysed in paper VI. The results from paper VI are extended in paper VII to address multi-view plus depth sequences of more than three views. A view priority and a depth priority strategy to the layer assignment of adjacent views are proposed and analysed. The assignment of the depth data to the corresponding layer of the color data is also addressed. The analysis of the layer assignment schemes in papers VI and VII was performed using objective metrics in relation to the rate of the sequence and complemented with visual examination. Each of the three papers are described in more detail in the rest of this section.

4.3.1 Paper V

The relation of the depth of a scene and the depth range of the display is used as a basis for a scalable coding scheme in [109]. The scheme proposed in paper V aims to divide the color part of the MVD data into enhancement layers that corresponds to the distance to the camera. This facilitates the extraction of the front most parts of each view. The approach is combined with view scalability [22] to ensure that the background objects can be rendered using data from the center view if no other data is available. This increases the quality of the center views.

The proposed depth and view scalability scheme is applied to a MVD sequence containing three views. The central view and depth data are assigned to the base layer. This ensures that the scheme never performs in a worse manner in the rendering step than a 2D plus depth video sequence. The two views to either side of the central view are divided into a limited number of enhancement layers in the encoding.

The decision to which enhancement layer a macro-block (MB) is assigned is based on two criteria. Firstly, the front most objects should be included in the first layer.

Secondly, the remaining pixels should be equally divided over the remaining enhancement layers. The statistics regarding the depth data are used to partition the depth domain into intervals that correspond to the two criteria. The first enhancement layer is determined in a similar manner to that for the first ROI detection approach in paper IV. The exception is that all MBs that contain pixels within this depth interval are included in the first layer. The remaining intervals are calculated from the cumulative distribution function of the depth data in order to distribute the number of pixel uniformly.

Quantitative tests were performed on an MVC encoder, modified to provide the proposed depth scalability in the two side views. The quality analysis, using PSNR as a metric, was applied to the color data after the rendering. The two decoded views closest to the virtual view were used in the rendering of each view. The original color data was used as a reference. The results showed that the addition of each enhancement layer provides an increased quality as measured both in PSNR and judged by visual appearance.

4.3.2 Paper VI

The approach proposed in paper V in the previous section divides the color data of an MVD video sequence into enhancement layers based on a uniform distribution of the pixels. The problem with a uniform distribution is that artifacts may appear in the rendering when the boundary of a layer intersects an object. Paper VI addresses two alternative schemes to assign color data to enhancement layers.

In both the alternative schemes, Scheme B and Scheme C, the depth interval of enhancement layer one is chosen based on the characteristics of the depth data distribution. Consequently, the position of the first layer can be adapted to the actual position of the front most objects. The assignment of pixels to the remaining layers differs in the two schemes. Scheme B applies the uniform distribution scheme of paper VI to determine the position of the remaining enhancement layers. In Scheme C the characteristics of the depth data distribution are used both to determine the position in depth of the remaining enhancement layers and the number of layers.

The three layer assignment schemes were evaluated concerning their effect on the rate and quality measured in PSNR of the sequence, PSNR per layer and visual appearance. As in paper V, the PSNR of the rendered virtual views were used. However, the reference data was rendered as described by figure 2.12, which reduces the effects of the rendering on the results. The complete sequence at a lower quantization parameter provides a better performance in the terms of overall rate-distortion. Then again, the PSNR per layer and the visual examples showed that the quality of the front-most objects are better for a sequence containing just a few enhancement

layers compared to the full sequence encoded at a larger quantization parameter.

The rate-distortion evaluation indicates that that a large number of enhancement layers increases the flexibility in the choice of bit rate. However, the chance of rendering artifacts due to the position of the layer border is increased and the coding efficiency is slightly reduced. The result concerning Scheme C showed that 4-6 layers were chosen using the characteristics of the depth distribution. This implies both that the ideal choice of layers depends on the video sequence and that the number should be limited.

4.3.3 Paper VII

The approach proposed in paper VI in the previous section is restricted to an MVD sequence of three views and only considers the depth scalability of the color data. Two strategies with regards as to how to extend the three view scheme to an arbitrary number of views are presented in paper VII. These strategies consider the relation between the layers of adjacent views that contain enhancement layers. The assumption is that the same scheme is used to assign the layers within each view. Should the scalability between views, or in depth, be given the higher priority? In the first proposed strategy, view priority, the view scalability is given the higher priority by extracting all the enhancement layers of one view before the next. The second proposed strategy, depth priority, gives a higher priority to layers containing data close to the camera regardless of the camera position.

The two strategies were tested using MVD sequences containing five views that were compressed using a modified MVC encoder. The layer assignment of individual views was made using Scheme C from paper VI, which has been renamed as depth distribution layer assignment (DLA). The quality was analyzed using the PSNR metric, as in paper VI. The PSNR per view and PSNR per depth metrics were also applied.

View priority gave the best result considering average PSNR, when layers are discarded. On the other hand, the visual quality evaluation indicated that depth priority gave a better visual quality. The difference between the measured result and the visual evaluation depends on the fact that the average PSNR metric does not show the improved quality of the frontal objects. In addition, the rendering in depth priority uses data from a view farther away than view priority. Some of the errors that are introduced by the rendering are not visible in the visual evaluation. This explains the difference in the result in PSNR and visual evaluation. The PSNR per view and PSNR per depth value confirm that depth priority generally performs better when only a few enhancement layers are included. However, view priority is the better choice if the quality of the center views is important in the application.

Another issue addressed in this paper is the encoding of the depth data using scalability. In the approach described in papers V and VI, all the depth data of a view is assigned to the first enhancement layer of that view. The scalable coding of the color data in this approach, can be extended to encompass the depth data. The depth data is then assigned to the same enhancement layers as the corresponding color data. Tests showed that the bit-rate of each added layer decreases with no noticeable effect on the PSNR or visual quality, when the depth data is distributed over the enhancement layers of a view compared to assigning all the depth data to the first enhancement layer. The cost is a slight decrease in coding efficiency for the full sequence.

Chapter 5

Conclusions

5.1 ROI video coding

Region of interest (ROI) video coding makes it possible to adapt the encoding with regards to how a human would perceive the quality of a video sequence in a particular application at low bit rates. The quality of the ROI can be improved by reducing the quality in the less noticeable background, which gives the appearance of improved perceived quality to the viewer without having to increase the bit rate.

The first goal adressed in this thesis was to propose a codec independent region-of-interest approach that is applied in both the temporal and spatial domain to increase the perceived quality in a 2D video sequence at a fixed bit rate. A pre-processing approach in the form of a spatio-temporal (SPTP) filter is proposed to achieve this goal. The filter contains a spatial (SP) part that removes details in the background and the temporal (TP) part which removes information in the background of every second frame.

The SP part reduces the number of bits that are allocated to the DCT components, the prediction error and the motion vectors of the background, due to the reduced prediction error. Multiple Gaussian filters are used to enable a gradual quality transition from ROI to background. Hence, artifacts at the ROI border due to a low cut-off frequency are avoided. The TP part, on the other hand, reduces the number of bits used for the motion vectors by reducing the number of motion vectors of the background. Bilinear interpolation can be applied to reduce artifacts at the border due to movement of the ROI.

In order for the pre-filtering to result in a successful ROI video coding, the bits released by the filter must be re-allocated to the ROI by the encoder. The encoder

assigns bits to where they have the greatest effect on the reduction of the distortion. The distortion by the number of DCT components (intra-frames) and the prediction error (inter-frames). Hence, the SP part is the key to a successful reallocation.

The SPTP filter combines the effect of the SP and TP parts to reduce the bits assigned to the background and successfully re-allocating these bits to the ROI. In addition, the SPTP filter has a lower computational complexity than the SP filter. The subjective test confirmed the result achieved using an objective metric that the SPTP filter improves the perceived quality at a fixed bitrate. Hence the first goal of this thesis is fulfilled.

The choice of the variances of the low-pass filters included in the SP filter has an impact on the coding efficiency of the SPTP filter. The larger the variance, the more of the DCT components that will be partially or completely suppressed and the higher the risk of the background distortion attracting attention from the ROI. The relation between the variance and the number of DCT components can be used to adapt the filter to any knowledge regarding the channel characteristics.

ROI video coding can also be extended to the color data of a 2D plus depth video sequence. The second goal of the thesis is to provide a scheme that extracts ROIs from this type of a sequence. Two detection schemes was proposed and analysed. In the first the additional depth data in 2D plus depth video was used to detect the ROIs with the assumption that objects close to the camera are of interest. The second scheme addressed the context of the video sequence by combining the first scheme with skin-color detection. It was shown in the visual examination that the content of the in objects in the background and the context in which they are shown may influence the perceived quality.

5.2 Multi-view plus depth scalable video coding

Consider the case of several receivers accessing the same 3D video content over a heterogeneous network. The problem is then how to provide the highest perceived quality for all end-users in this network. A solution is scalable video coding that enables the extraction of a part of a 3D video sequence. Hence, the video can be adapted to the limitations of a part of the network and/or an end-user. Scalable coding schemes for 2D video and known view scalability schemes can be applied to multi-view plus depth (MVD) video.

The third and last goal of this thesis was to propose a scalable coding algorithm of MVD data that addresses the depth domain using the concept that some regions in the video sequence are more important than others. A depth scalability scheme is proposed that divides the data within a view depending on the location of the con-

tent in order to achieve the third goal. The scheme is combined with view scalability to ensure that background content can be rendered using at least the center view data.

The proposed depth and view scalability scheme is applied on an MVD sequence, where the central view and corresponding depth data is assigned to the base layer. Each of the remaining views is divided into enhancement layers. Hence, it is possible to extract a subset of the pixels of a video sequence in the case of limited resources. The quality analysis showed that the quality of the front-most objects is better for a sequence containing just a few enhancement layers compared to the full sequence encoded with a larger quantization parameter.

Three types of schemes are proposed that assign the color data to a particular enhancement layer. These are all based on the distribution of depth data. The schemes either distribute the pixels uniformly over the enhancement layers, adapt the position of the layers to the depth distribution or a combination of the two. The problem associated with assigning layers by the uniform distribution of the depth data is that the layer border may intersect with objects resulting in artifacts in the rendering. A solution to the problem is the depth distribution layer assignment (DLA) scheme, which uses the characteristics of the depth data distribution to control both the number and the position of the enhancement layers. The bit rate of each layer can be reduced further if the depth data is divided into layers, in a similar manner to that for the color data. This involves a small cost namely a reduced coding efficiency for the whole sequence.

The layer assignment between views must be taken into account when the MVD sequence contains more than three views. Should depth or view scalability have the higher priority? The tests showed that the best visual quality over all views is achieved by extracting the front-most layers in all of the view before layers that contain data further back in the scene. The exception involves applications where the quality of the center views is considered to be the more important.

5.3 Future works

5.3.1 ROI video coding

Future works on ROI video coding can include:

- to improve the detection of ROI in both 2D and 3D video. Even though face detection has been extensively researched, there is still a need for faster and more accurate methods. In addition other applications, such as surveillance

and soccer games, could be considered, where methods of detection have not been thoroughly researched at the present time.

- to integrate the SP part of the SPTP filter into the codec and adapt the rate-distortion optimization to the low pass filter.
- to further address the problem associated with a moving ROI for the temporal filter. Perhaps it might be possible to include an optional post-processing step to reduce the impact of jerky movements in the background. This post-processing would attempt to recreate the information in the background that was removed by the temporal filter.
- to extend the codec-independent ROI coding scheme for 2D plus depth video to multi-view plus depth video. ROI coding of MVD has been addressed by Zhang et al [112], which uses codec-dependent ROI coding and disregarding interview prediction by using simulcast.

5.3.2 Multi-view plus depth scalable video coding

Future works on multi-view plus depth scalable vide coding can include:

- to combine the assignment of enhancement layers in the proposed scheme with rate distortion optimization. The bit rate reduction for each enhancement layer that is removed can then be evenly distributed over the layers.
- to combine the proposed depth and view scalability approach with the various scalability methods in the SVC amendment to H.264 AVC and analyze their combined behavior. The test result could be used to make a priority between the different types of scalability. The information could also be used to extend the SVC approach in [113] to 3D video. The scheme proposed by Akyol et al in [113] adapts the scalability type to the 2D video content.
- to evaluate the the proposed method and other SVC algorithms using a subjective test with a variety systems for visualizing 3D video.

Bibliography

[1] J.-Y. Son, B. Javidi, and K.-D. Kwack. Methods for displaying three-dimensional images. In *Proceedings of IEEE*, volume 94, pages 502–523, 2006.

[2] Moving Pictures Experts Group. http://www.chiariglione.org/mpeg/.

[3] International Telecommunications Union. http://www.itu.int/home/.

[4] ISO/IEC 13818-2. Information technology - generic coding of moving pictures and associated audio information: Video. International Organization for Standardization, 1996.

[5] ITU-T Rec. H.261. Video codec for audiovisual services at px64 kbit/s. International Telecommunications Union, 1993.

[6] ITU-T Rec. H.263. Video coding for low bitrate communication. International Telecommunication Union, 1998.

[7] ISO/IEC 14496-2. Information technology - coding of audio-visual objects part 2: Visual. International Organization for Standardization, 1999.

[8] ITU-T Rec. H.264. Advanced video coding for generic audiovisual services. International Telecommunication Union, 2003.

[9] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13:560–576, July 2003.

[10] ISO/IEC 23003-3. Mpeg-c part 3: Representation of auxiliary video and supplemental information. October 2007.

[11] H. Schwartz, D. Marple, and T. Wiegand. Overview of the Scalable Video Coding Extension of H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1103–1120, September 2007.

[12] Y. Wang, J. Ostermann, and Y.-Q. Zhang. *Video Processing and Communications*. Prentice-Hall Inc., Upper Saddle River, New Jersey, USA, 2002.

[13] E. A. Umble. Making it real: The future of stereoscopic 3D film technology. ACM Siggraph Computer Graphics, volume 40, 2006.

[14] SmartHouse: The lifestyle technology guide. Sony Vs Panasonic 3D HD TV Battle Emerges. http://www.smarthouse.com.au/TVs_And_Large_Display/3D_TV/D5C4Q7U4, 2009-09-29.

[15] J. Flack, J. Harrold, and G.J. Woodgate. A Prototype 3D Mobile Phone Equipped with a Next Generation Autostereoscopic Display. In *SPIE*, volume 6490A-21, pages 502–523, 2007.

[16] L. Lewis. Sony triumphs with 3D gaming technology. http://technology.timesonline.co.uk, 2009-09-24.

[17] C. Chinnock. 3D Comming Home in 2010. http://www.3dathome.org, October 2009.

[18] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang. Multiview Imaging and 3DTV. *IEEE Signal Processing Magazine*, 24:10 – 21, November 2007.

[19] A. Smolic, K. Müller, P. Merkle, P. Kauff, and T. Wiegand. An overview of avaliable and emerging 3d video formats and depth enhanced stereo as efficient generic solution. In *Picture Coding Symposium*. IEEE, 2009.

[20] 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*, 17:1461–1473, November 2007.

[21] P. Merkle, K. Müller, and T. Wiegand. Efficient compression of multiview video exploiting inter-view dependencies based on H.264/MPEG4-AVC. In *ICME*, pages 1717–1720. IEEE, 2006.

[22] J.E. Lim, K.N. Ngan, W. Yang, and K. Sohn. A multiview sequence CODEC with view scalability. *Signal Processing: Image Communication*, 19:239–256, 2004.

[23] M. Flierl and B. Girod. Multiview video compression. *IEEE Signal Processing Magazine*, 24:66–76, Nov. 2007.

[24] J. Lu, H. Cai, J.-G. Lou, and J. Li. An epipolar geometry-based fast disparity estimation algorithm for multiview image and video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:737–750, June 2007.

[25] U. Fecker, M. Barkowsky, and A. Kaup. Improving the prediction efficiency for multi-view video coding using histogram matching. In *Picture Coding Symposium, Beijing, China*, 2006.

[26] J.-H. Hur, S. Cho, and Y.-L. Lee. Adaptive Local Illumination Change Compensation Method for H.264/AVC-Based Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1496 – 1505, November 2007.

[27] J.-H. Kim, P. Lai, J. Lopez, A. Ortega, Y. Sun, P. Yin, and C. Gomila. New coding tools for illumination and focus mismatch compensation in multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1519 – 1535, November 2007.

[28] K. Müller, P. Merkle, and T. Wiegand. Compressing time varying visual content. *IEEE Signal Processing Magazine*, 24:58–65, Nov. 2007.

[29] C. Fehn. Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In *Stereoscopic Displays and Virtual Reality Systems XI*, pages 93–104. SPIE, 2004.

[30] P. Merkle, A. Smolic, K. Müller, and T. Wiegand. Multi-view video plus depth representation and coding. In *ICIP*, volume I, pages 201–205, 2007.

[31] I. Daribo, C. Tillier, and B. Pesquet-Popescu. Motion Vector Sharing and Bitrate Allocation for 3D Video-Plus-Depth Coding. *EURASIP Journal on Advances in Signal Processing*, 2009.

[32] M.T. Pourazad, P. Nasipoulos, and R.K. Ward. An H.264 based video encoding scheme for 3D TV. In *EUSIPCO*. EURASIP, 2006.

[33] R. Krishnamurthy, B-B. Chai, H. Tao, and S. Sethuraman. Compression and transmission of depth maps for image based rendering. In *2001 International Conference on Image Processing , ICIP, Thessaloniki, Greece*, 2001.

[34] P. Zanuttigh and G.M. Cortelazzo. Compression of depth information for 3D rendering. In *3DTV-CON*. IEEE, 2009.

[35] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szelinski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23:600 – 608, August 2004.

[36] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P.H.N de With, and T. Wiegand. The effect of depth compression on multiview rendering quality. In *3DTV-CON*. IEEE, 2008.

[37] E. Ekmekcioglu, S. Worrall, and A.M. Kondoz. A temporal subsampling approach for multiview depth map compression. *IEEE Transactions on circuits and systems for video technology*, 19:1209–1213, August 2009.

[38] Y.K. Park, K. Jung, Y. Oh, S. Lee, J.K. Kim, G. Lee, H. Lee, K. Yun, N. Hur, and J. Kim. Depth-image-based rendering for 3DTV services over T-DMB. *Signal Processing: Image Communication*, 24:122–139, 2009.

[39] O. Schreer, P. Kauff, and T. Sikora. *3D Videocommunication: Algorithms, concepts and real-time systems in human centered communication*. John Wiley & Sons Ltd, Chichester, England, 2005.

[40] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. View Synthesis for Advanced 3D Video Systems. *EURASIP Journal on Image and Video Processing*, 2008.

[41] E. Cooke, P. Kauff, and T. Sikora. Multiview synthesis: A novel view creation approach for free viewpoint video. *Signal Processing: Image Communication*, 21:476–492, 2006.

[42] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto. View generation with 3D warping using depth information for FTV. *Signal Processing: Image Communication*, 24:65–72, 2009.

[43] A. Telea. An image inpainting tehcnique based on the fast marching method. *Journal of Graphics Tools*, 9:25–36, 2004.

[44] S.H. Kang, T.F. Chan, and S. Soatto. Inpainting from Multiple Views. In *First International Symposium on 3D Data Processing Visualization and Transmission, 3DPVT'02*, 2002.

[45] K.J Oh, S. Yea, and Y-S. Ho. Hole-Filling Method Using Depth Based In-Painting For View Synthesis in Free Viewpoint Television (FTV) and 3D Video. In *Picture Coding Symposium*. IEEE, 2009.

[46] Z-W. Liu, P. An, S-X. Liu, and Z-Y. Zhang. Arbitrary view generation based on DIBR. In *2007 International Symposium on Intelligent Signal Processing and Communication Systems, Xiamen, China*, pages 168–171, 2007.

[47] I. Daribo, C. Tillier, and B. Pesquet-Popescu. Distance Dependent Depth Filtering in 3D Warping for 3DTV. In *2007 IEEE 9th Workshop on Multimedia Signal Processing, MMSP, Chania, Greece*, pages 312–315, 2007.

[48] L. Zhang and W.J. Tam. Stereoscopic image generation based on depth images for 3DTV. *IEEE Transactions on Broadcasting*, 51:191–199, 2005.

[49] Y. Huang and C. Zhang. A layered method of visibility resolving in depth image-based rendering. In *International Conference on Pattern Recognition*. IEEE, 2008.

[50] ANSI T1.80103-1996. Digital transport of one-way video signals parameters for objective performance assessment. American National Standards Institute, 1996.

[51] Z. Wang, H.R. Sheikh, and A.C. Bovik. Chapter 41 in the handbook of video databases: Design and application. B. Furth and O. Marqure, CRC Press, pp 1041-1078, 2003.

[52] S. Lee, M.S. Pattichis, and A.C. Bovik. Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4:129–132, March 2002.

[53] C.J. van den Branden Lambrecht, O. Versheure, J. Urbain, and F. Tassin. Perceptual quality measure using a spatio-temporal model of the human visual system. In *SPIE*, volume 2668, pages 450–461, 1996.

[54] Z. Wang, A.C. Bovik, and L. Lu. Why is image quality assessment so difficult. In *2002 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 3313–3317, 2002.

[55] M.H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, pages 312 – 322, September 2004.

[56] R. Olsson and M. Sjöstrom. A depth dependent quality metric for evaluation of coded integral imaging based 3D-images. In *ICIP*, pages 513–516. IEEE, 2006.

[57] N. Ozbek and A.M. Tekalp. Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video measure. In *International Conference on Multimedia and Expo*, pages 1113–1116. IEEE, 2008.

[58] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau. Quality assesment of stereoscopic images. *EURASIP Journal on Image and Video Processing*, 2008.

[59] ITU-R Rec. BT.500-11. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, 2002.

[60] ITU-R Rec. BT.1438. Subjective assessment of stereoscopic television pictures. International Telecommunication Union, 2000.

[61] A. Cavender, R. Vanam, D.K. Barney, R.E Ladner, and E.A Riskin. MobileASL: Intelligibility of sign language video over mobile phones. *Disabilility and Rehabilitation:Assistive Technology*, 3:93–105, January 2008.

[62] A. Eleftheriadis and A. Jacquin. Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates. *Signal Processing: Image Communication*, 7:231–248, September 1995.

[63] M.-J. Chen, M.-C. Chi, C.-T. Hsu, and J.-W. Chen. ROI Video Coding Based on H.263+ with Robust Skin-Color Detection Technique. *IEEE Transactions on Consumer Electronics*, 49:724–730, August 2003.

[64] L. Karlsson. Detection of interesting areas in images by using convexity and rotational symmetries. Master Thesis No. 31 (2002), Dept. of Science and Technology, Linköping University, Sweden, 2002.

[65] J. Meessen, C. Parisot, X. Desurmont, and J.-F. Delaigle. Scene Analysis for Reducing Motion JPEG 2000 Video Surveillance Delivery Bandwidth and Complexity. In *2005 IEEE International Conference on Image Processing, Genua, Italy*, volume 1, pages 577–580, 2005.

[66] G. Wang, T.-T. Wong, and P.-A. Heng. Real-Time Surveillance Video Display with Salience. In *The Third ACM International Workshop on Video Surveillance & Sensor Networks*, pages 37–43, 2005.

[67] J.D. McCarthy, M.A. Sasse, and D. Miras. Sharp or Smooth? Comparing the effects of quantization vs. frame rate for streamed video. In *The SIGCHI Conference on Human Factors in Computing Systems*, pages 535–542, 2004.

[68] Y.-L. Kang, J.-H. Lim, Q. Tian, M.S. Kankanhalli, and C.-S. Xu. Visual Keywords Labeling in Soccer Video. In *The 17th International Conference on Pattern Recognition*, pages 535–542, 2004.

[69] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Human Neurobiology, Springer-Verlag*, pages 219–227, 1985.

[70] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1254–1259, November 1998.

[71] M. Bollmann, R. Hoischen, and B. Mertsching. Integration of static and dynamic scene features guiding visual attention. In *Mustererkennung 1997, 19th DAGM-Symposium*, pages 483–490, 1997.

[72] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu. A User-Attention Based Focus Detection Framework and Its Applications. In *The 4th International Conference on Informations, Communications and Signal Processing and Fourth Pacific-Rim Conference on Multimedia*, volume 3, pages 1315–1319, 2003.

[73] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel. Task Impact on the Visual Attention in Subjective Image Quality Assessment. In *The 14th European Signal Processing Conference*, 2006.

[74] M.-H. Yang, D.J. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:34–58, January 2002.

[75] V. Vezhnevets, V. Sazonov, and A. Andreeva. A Survey on Pixel-Based Skin Color Detection Techniques. In *Graphicon-2003, Moscow, Russia*, pages 85–92, 2003.

[76] J. Ahlberg. A system for face localization and facial feature extraction. Technical report, Department of Electrical Engineering, Linköping University, Linöping, Sweden, 1999.

[77] Y.H Chan and S.A.R Abu-Bakar. Face Detection System Based on Feature-Based Chrominance Colour Information. In *The International Conference on Computer Graphics, Imaging and Visualization*, pages 153–158, 2004.

[78] Y.-X. Lv, Z.-Q. Liu, and X.-H. Zhu. Real-Time Face Detection based on Skin-Color Model and Morphological filters. In *The 2nd International Conference on Machine Learning and Cybernetics*, pages 3203–3207, 2003.

[79] J.-C. Terrillon, M.N Shiraz, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *The International Conference on Face and Gesture Recognition*, pages 54–61, 2000.

[80] L.-H. Zhao, X.-L. Sun, J.-H. Liu, and X.-H. Xu. Face Detection based on Skin Color. In *The 3rd International Conference on Machine Learning and Cybernetics*, pages 3625–3628, 2004.

[81] M.-H. Yang and N. Ahuja. Gaussian mixture model for human skin color and its application in image and video databases. In *The SPIE:Conference on Storage and Retrival for Image and Video Databases*, pages 458–466, 1999.

[82] M.J. Jones and J.M. Rehg. Statistical Color Models with Application to Skin Detection. In *The International Conference on Computer Vision and Pattern Recognition*, pages 274–280, 1999.

[83] S.L. Phung, D. Chai, and A. Bouzerdoum. Adaptive Skin Segmentation in Color Images. In *The 3rd International Conference on Machine Learning and Cybernetics*, pages 3625–3628, 2004.

[84] R.-L. Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face Detection in Color Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:696–706, May 2002.

[85] J. Ruiz del Solar and R. Vershae. Skin Detection using Neighborhood Information. In *The 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[86] J.-B. Lee and A. Eleftheriadis. Spatio-Temporal Model-Assisted Very Low-Bit-Rate Coding With Compatibility. *IEEE Transactions on Circuits and Systems for Video Technology*, 15:1517–1531, December 2005.

[87] D. Brown, I. Craw, and J. Lewthwaite. A SOM Based Approach to Skin Detection with Application in Real Time Systems. In *The British Machine Vision Conference*, 2001.

[88] D. Chai, K.N. Ngan, and A. Bouzerdoum. Foreground/Background Bit Allocation for Region-of-Interest Coding. In *2000 IEEE International Conference on Image Processing*, volume 2, pages 923–926, 2000.

[89] S. Daly, K. Matthews, and J. Ribas-Corbera. Face-Based Visually-Optimized Image Sequence Coding. In *1998 IEEE International Conference on Image Processing*, volume 3, pages 443–447, 1998.

[90] S. Sengupta, S.K. Gupta, and J.M. Hannah. Percetually Motivated Bit-Allocation for H.264 Encoded Video Sequences. In *2003 IEEE International Conference on Image Processing*, volume 3, pages 797–800, 2003.

[91] X.K. Yang, W.S. Lin, Z.K. Lu, X. Lin, S. Rahrdja, E.P. Ong, and S.S Yao. Local Visual Perceptual Clues and its use in Videophone Rate Control. In *The 2004 International Symposium on Circuits and Systems*, volume 3, pages 805–808, 2004.

[92] T. Adiono, T. Isshiki, K. Ito, and T. Ohtsuka. Face Focus Coding under H.263+ Video Coding Standard. In *IEEE Asia-Pacific Conference on Circuits and Systems*, volume 3, pages 461–464, 2000.

[93] H. Wang and K. El-Maleh. Joint Adaptive Background Skipping and Weighted Bit Allocation for Wireless Video Telephony. In *2005 IEEE International Conference on Wireless Networks, Communications and Mobile Computing*, pages 1243–1248, 2005.

[94] Y. Liu and K. Ngi Ngan. Fully scalable multiview wavelet video coding. In *International Symposium on Circuits and Systems*, pages 2581 – 2584. IEEE, 2009.

[95] P. Sivanantharasa, W.A.C. Fernando, and H. Kodikara Arachchi. Region of interest video coding with flexible macroblock ordering. In *International Conference on Industrial and Information Systems*, pages 596–599. IEEE, 2006.

[96] L. Itti. Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. *IEEE Transactions on Image Processing*, 13:1304–1318, October 2004.

[97] J.-W. Lee, A. Vetro, Y. Wang, and Y.-S. Ho. Bit Allocation for MPEG-4 Video Coding With Spatio-Temporal Tradeoffs. *IEEE Transactions on Circuits and Systems for Video Technology*, 13:488–502, June 2003.

[98] J. Augustine, S.K. Rao, N.P. Jouppi, and S. Iyer. Region of Interest Editing of MPEG-2 Video Streams in the Compressed Domain. In *2004 IEEE International Conference on Multimedia and Expo*, pages 559–562, 2004.

[99] H. Wang, Y. Liang, and K. El-Maleh. Real-Time Region-of-Interest Video Coding using Content-Adaptive Background Skipping with Dynamic Bit Reallocation. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 45–48, 2006.

[100] G. Cheung, T. Sakamoto, and W. Tan. Graphics-to-video encoding for 3G mobile game viewer multicast using depth values. In *International Conference on Image Processing*, volume 4, pages 2805–2808. IEEE, 2004.

[101] E. Masala and D. Quaglia. Perceptually optimized mpeg compression of synthetic video sequences. In *ICIP*, volume 1, pages 601–604, 2005.

[102] H. Aydinoglu, F. Kossentini, Q. Jiang, and M.H. Hayes. Region-based stereo image coding. In *International Conference on Image Processing*, pages 57–60. IEEE, 1995.

[103] O. Marques, L.M. Mayron, D. Socek, G.B. Borba, and H.R. Gamba. An attention-based method for extracting salient regions of interest from stereo images. In *International Conference on Computer Vision Theory and Applications*, pages 143–148, 2007.

[104] L.S Karlsson, R. Olsson, and M. Sjöström. Temporal filter with bilinear interpolation for ROI video coding. Technical report, MUCOM, Department of Information Technology and Media, Mid Sweden University, Sundsvall, Sweden, http://miun.diva-portal.org/smash/record.jsf?searchId=1&pid=diva2:30912, 2006.

[105] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1238 – 1255, September 2007.

[106] M. Dröse, C. Clemens, and T. Sikora. Extending single-view scalable video coding to multi-view base on h.264/AVC. In *ICIP*, pages 2977–2980, 2006.

[107] J. Cho, S. Cho, N. Hur, H. Lee, and J. Jeong. Effective multiview video coding using a scalable depth map. In *International Conferences on Computational Intelligence for Modelling, Control and Automation*, pages 255 – 259. IEEE, 2008.

[108] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima. View Scalable Multiview Video Coding Using 3-D Warping With Depth Map. *IEEE Transactions on Circuits and Systems for Video Technology*, 17:1485–1495, November 2007.

[109] V. Ramachandra, M. Zwicker, and T.Q. Nguyen. Display dependent coding for 3D video on automultiscopic displays. In *ICIP*, pages 2436–2439, 2008.

[110] W. Yang, F. Wu, J. Cai, K. Ngi Ngan, and S. Li. Scalable multiview video coding using wavelets. In *International Symposium on Circuits and Systems*, pages 6078 – 6081. IEEE, 2005.

[111] J-U. Garbas, U. Fecker, T. Troger, and A. Kaup. 4D Scalable Multi-View Video Coding Using Disparity View Filtering and Motion Compensated Temporal Filtering . In *Workshop on Multimedia Signal Processing*, pages 54 – 58. IEEE, 2006.

[112] Y. Zhang, M. Yu, and G.Y. Jiang. Depth based region of interest extraction for multi-view video coding. In *Eight International Conference on Machine Learning and Cybernetics , Baoding, China*, pages 2221–2226, 2009.

[113] E. Akyol, A.M. Tekalp, and M.R. Civanlar. Content-Aware Scalability-Type Selection for Rate Adaption of Scalable Video. *EURASIP Journal on Advanced Signal Processing*, 2007.

Biography

Linda S. Karlsson was born on the 4th of April 1976 in Katrineholm, Sweden. She received the Master of Science degree in Media technology and and engineering from Linköping University, Sweden in September 2002 and an the degree Licenciate of Technology in Computer and System Sciences from Mid Sweden University, Sweden in May 2007. During the studies before the university she had the opportunity to enroll as a student at Robert Johnson High School in Gainesville, Georgia in the USA in 1993–1994. She has also attended a year of exchangestudies at the Rheinisch-Westfälische Technische Hochschule in Aachen, Germany in 2000–2001.

Karlsson is currently a PhD Student at the department of Information Technology and Media at Mid Sweden University. Her main research interest are in the field of video analysis and video source coding. Currently she is working on standard and codec independent region-of-interest video coding and multiview-plus-depth scalable video coding.