



TAMPEREEN TEKNILLINEN YLIOPISTO  
TAMPERE UNIVERSITY OF TECHNOLOGY

Payman Aflaki Beni

**Compression and Subjective Quality Assessment of  
3D Video**



Julkaisu 1174 • Publication 1174

Tampere 2013

Tampereen teknillinen yliopisto. Julkaisu 1174  
Tampere University of Technology. Publication 1174

Payman Aflaki Beni

## **Compression and Subjective Quality Assessment of 3D Video**

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 29<sup>th</sup> of November 2013, at 12 noon.

ISBN 978-952-15-3184-2 (printed)  
ISBN 978-952-15-3213-9 (PDF)  
ISSN 1459-2045

# Abstract

In recent years, three-dimensional television (3D TV) has been broadly considered as the successor to the existing traditional two-dimensional television (2D TV) sets. With its capability of offering a dynamic and immersive experience, 3D video (3DV) is expected to expand conventional video in several applications in the near future. However, 3D content requires more than a single view to deliver the depth sensation to the viewers and this, inevitably, increases the bitrate compared to the corresponding 2D content. This need drives the research trend in video compression field towards more advanced and more efficient algorithms.

Currently, the Advanced Video Coding (H.264/AVC) is the state-of-the-art video coding standard which has been developed by the Joint Video Team of ISO/IEC MPEG and ITU-T VCEG. This codec has been widely adopted in various applications and products such as TV broadcasting, video conferencing, mobile TV, and blue-ray disc. One important extension of H.264/AVC, namely Multiview Video Coding (MVC) was an attempt to multiple view compression by taking into consideration the inter-view dependency between different views of the same scene. This codec H.264/AVC with its MVC extension (H.264/MVC) can be used for encoding either conventional stereoscopic video, including only two views, or multiview video, including more than two views.

In spite of the high performance of H.264/MVC, a typical multiview video sequence requires a huge amount of storage space, which is proportional to the number of offered views. The available views are still limited and the research has been devoted to synthesizing an arbitrary number of views using the multiview video and depth map (MVD). This process is mandatory for auto-stereoscopic displays (ASDs) where many views are required at the viewer side and there is no way to transmit such a relatively huge number of views with currently available broadcasting technology. Therefore, to satisfy the growing hunger for 3D related applications, it is mandatory to further decrease the bitstream by introducing new and more efficient algorithms for compressing multiview video and depth maps.

This thesis tackles the 3D content compression targeting different formats i.e. stereoscopic video and depth-enhanced multiview video. Stereoscopic video compression algorithms introduced in this thesis mostly focus on proposing different types of asymmetry between the left and right views. This means reducing the quality of one view compared to the other view aiming to achieve a better subjective quality against the symmetric case (the reference) and under the same bitrate

constraint. The proposed algorithms to optimize depth-enhanced multiview video compression include both texture compression schemes as well as depth map coding tools. Some of the introduced coding schemes proposed for this format include asymmetric quality between the views.

Knowing that objective metrics are not able to accurately estimate the subjective quality of stereoscopic content, it is suggested to perform subjective quality assessment to evaluate different codecs. Moreover, when the concept of asymmetry is introduced, the Human Visual System (HVS) performs a fusion process which is not completely understood. Therefore, another important aspect of this thesis is conducting several subjective tests and reporting the subjective ratings to evaluate the perceived quality of the proposed coded content against the references. Statistical analysis is carried out in the thesis to assess the validity of the subjective ratings and determine the best performing test cases.

# Acknowledgments

The research work included in this thesis covers a total four-year time research during years 2009-2013 which I was working in Tampere University of Technology (TUT) and Nokia Research Center (NRC). In TUT I was working as a researcher in Multimedia Research Group at the Department of Signal Processing and at NRC as an external researcher in Multimedia Visual Technologies group.

This thesis owes its existence to the help and inspiration of several people. First and foremost, my sincere gratitude goes to my supervisor, Prof. Moncef Gabbouj whose guidance and encouragement supported me continuously throughout my PhD studies. I would like to also thank Prof. Gabbouj for enabling collaboration between TUT and NRC to better familiarize me with industrial requirements as well as academic research.

I am indebted to Dr. Miska Hannuksela for his constant technical supervision on every detail of my research during these four years and inspiring me to perform efficiently. Without having him illuminating the path for me, this thesis could not be accomplished.

I would like to thank the reviewers of this thesis, Prof. Lina Karam from the School of Electrical, Computer, and Energy Engineering at Arizona State University and Prof. Yao Wang from the Department of Electrical Engineering at Polytechnic Institute of New York University for their valuable feedback. I would also like to thank Dr. Dmytro Rusanovskyy for not only guiding me during the second half of my research, but also for motivating me towards a better future. Moreover, I would like to thank Virve Larmila, Ulla Siltaloppi, and Elina Orava who made all the administration regulation smooth and fast. My warm thanks go to my friends Alireza Razavi, Hamed Sarbolandi, Jenni Hukkanen, and Markus Penttila whose company made these years very memorable for me.

I also thank Nokia foundation and TUT for granting me several scholarships.

I would also like to thank Hadis Behzadifar for her full support, kindness, and patience during last years.

Last but not least, I would like to thank my parents, Mansour and Habibeh, from the bottom of my heart for their devotion to my success and persistent confidence in me. Words fail me to express my appreciation to my brother Aman, as he has been and will continue to act as the hero of my life. Finally, I dedicate this thesis to my parents and brother.

Payman Aflaki Beni

November, 2013



# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>List of Publications</b>	<b>vii</b>
<b>List of Abbreviations</b>	<b>ix</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objectives and outline of the thesis . . . . .	3
1.2 Publications and author's contribution . . . . .	5
<b>2 Human Visual System</b>	<b>7</b>
2.1 Binocular human vision . . . . .	9
2.2 Spatial perceptual information . . . . .	10
2.3 Binocular suppression theory . . . . .	14
<b>3 3D Content Visualization</b>	<b>15</b>
3.1 Scene characteristics . . . . .	15
3.2 3D displays . . . . .	16
3.3 Stereoscopic displays . . . . .	16
3.3.1 Passive displays . . . . .	17
3.3.2 Active displays . . . . .	17
3.4 Auto-stereoscopic displays . . . . .	18
3.4.1 Dual-view auto-stereoscopic displays . . . . .	19
3.4.2 Multiview auto-stereoscopic displays . . . . .	19



---

<b>4</b>	<b>Quality Assessment of 3D Video</b>	<b>23</b>
4.1	Objective metrics . . . . .	24
4.2	Subjective quality assessment . . . . .	29
4.2.1	Test procedure . . . . .	29
4.2.2	Analyzing subjective scores . . . . .	30
4.3	Subjective quality of 3D video . . . . .	31
4.3.1	Viewing 3D content with glasses . . . . .	31
4.3.2	Viewing 3D content without glasses . . . . .	32
<b>5</b>	<b>Asymmetric Stereoscopic Video</b>	<b>35</b>
5.1	Introduction . . . . .	35
5.2	Types of asymmetry . . . . .	37
5.3	Motivation for using asymmetric stereoscopic video . . . . .	38
5.3.1	Low-pass filtering . . . . .	38
5.3.2	Down/up sampling . . . . .	39
5.3.3	Performance analysis of different asymmetric types . . . . .	40
5.4	Limits of asymmetry . . . . .	47
5.5	Modeling subjective ratings . . . . .	48
5.6	Summary . . . . .	50
<b>6</b>	<b>Depth-Enhanced Multiview Video Compression</b>	<b>53</b>
6.1	Introduction . . . . .	53
6.2	Depth map . . . . .	55
6.3	Synthesizing virtual views . . . . .	56
6.4	Quality dependency of rendered views . . . . .	57
6.5	Depth map compression . . . . .	59
6.5.1	Depth map filtering . . . . .	60
6.5.2	Depth down/up sampling . . . . .	60
6.6	Using asymmetry in multiview video compression . . . . .	62
6.6.1	Asymmetric quality . . . . .	62
6.6.2	Mixed-resolution texture . . . . .	62
6.7	Video compression artifacts . . . . .	63
6.8	Summary of subjectively assessed experiments . . . . .	63
<b>7</b>	<b>Conclusion and Future Work</b>	<b>67</b>
7.1	Future work . . . . .	68
	<b>Bibliography</b>	<b>69</b>
	<b>Appendix - Publications</b>	<b>85</b>

# List of Publications

This thesis consists of the following publications.

- [P1] **P. Aflaki**, M. M. Hannuksela, D. Rusanovskyy, and M. Gabbouj, “Non-linear depth map resampling for depth-enhanced 3D video coding, ” *IEEE Signal Processing Letters*, Vol. 20, issue 1, pp. 87-90, January, 2013.
- [P2] **P. Aflaki**, M. M. Hannuksela, H. Sarbolandi, and M. Gabbouj, “Simultaneous 2D and 3D perception for stereoscopic displays based on polarized or active shutter glasses, ” *Elsevier Journal of Visual Communication and Image Representation*, March, 2013.
- [P3] **P. Aflaki**, M. M. Hannuksela, and M. Gabbouj; “Subjective quality assessment of asymmetric stereoscopic 3-D video, ” *Springer Journal of Signal, Image and Video Processing*, 2013.
- [P4] **P. Aflaki**, D. Rusanovskyy, M. M. Hannuksela, and M. Gabbouj; “Unpaired multiview video plus depth compression, ” *IEEE Digital Signal Processing*, Santorini, Greece, July, 2013.
- [P5] **P. Aflaki**, Wenyi Su, Michal Joachimiak, D. Rusanovskyy, M. M. Hannuksela, Houqiang Li, and M. Gabbouj; “Coding of mixed-resolution multiview video in 3D video application, ” *IEEE International Conference on Image Processing (ICIP)*, Melbourne, Australia, September, 2013.
- [P6] **P. Aflaki**, M. M. Hannuksela, M. Homayouni, and M. Gabbouj; “Cross-asymmetric mixed-resolution 3D video compression, ” *International 3DTV CONF*, Zurich, Switzerland, October. 2012.
- [P7] **P. Aflaki**, D. Rusanovskyy, M. M. Hannuksela, and M. Gabbouj; “Frequency based adaptive spatial resolution selection for 3D video coding, ” *European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August, 2012.
- [P8] **P. Aflaki**, D. Rusanovskyy, T. Utriainen, E. Pesonen, M. M. Hannuksela, S. Jumisko-Pyykkö, and M. Gabbouj; “Study of asymmetric quality between coded views in depth-enhanced multiview video coding, ” *International Conference on 3D Imaging (IC3D)*, Liege, Belgium, December, 2011.

- [P9] **P. Aflaki**, M. M. Hannuksela, J. Hakala, J. Häkkinen, M. Gabbouj; “Joint Adaptation of Spatial Resolution and Sample Value Quantization for Asymmetric Stereoscopic Video Compression: a Subjective Study, ” International Symposium on Image and Signal Processing and Analysis (ISPA) , Dubrovnik, Croatia, September 2011.
- [P10] **P. Aflaki**, M. M. Hannuksela, J. Hakala, J. Häkkinen, M. Gabbouj; “Estimation of subjective quality for mixed-resolution stereoscopic video, ” International 3DTV CONF. Antalya, Turkey, May 2011.

# List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
3D QA	3D Quality Assessment
3DV	3D Video
ASD	Auto-Stereoscopic Display
AVC	Advanced Video Coding
BVQM	Batch Video Quality Metric
CfP	Call for Proposals
CI	Confidence Interval
CSF	Contrast Sensitivity Function
CTC	Common Test Conditions
DCT	Discrete Cosine Domain
DIBR	Depth Image Based Rendering
DM	Distortion Measure
DSIS	Double Stimulus Impairment Scale
FR	Full-Resolution
FRef	Full-Reference
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IQA	Image Quality Assessment
LDV	Layered Depth Video
LERP	Linear Interpolation
LGN	Lateral Geniculate Nucleus
LPF	Low-Pass Filter
MR	Mixed-Resolution
MSE	Mean Square Error
MVC	Multiview Video Coding
MVD	Multiview Video plus Depth
NQM	Noise Quality Measure
NRef	No-Reference
PPD	Pixels Per Degree
PSF	Point Spread Function
PSNR	Peak-Signal-to-Noise
QP	Quantization Parameter

RDO	Rate-Distortion Optimization
RGB	Red, Green and Blue
RRef	Reduced-Reference
SAD	Sum of Absolute Differences
SD	Standard Definition
SI	Spatial Information
SLERP	Spherical Linear Interpolation
SSD	Statistical Significant Difference
SSIS	Single Stimulus Impairment Scale
TFT-LCD	Thin Film Transistor Liquid Crystal Display
TV	Television
UQI	Universal Quality Index
VQM	Video Quality Metric

# List of Figures

1.1	Different formats to present 3D content . . . . .	2
2.1	Cones and rods in the retina . . . . .	8
2.2	Left and right perspective of stereoscopic content . . . . .	9
2.3	Functional model of binocular vision . . . . .	11
2.4	Panum’s fusional areas . . . . .	12
3.1	Auto-stereoscopic display . . . . .	18
3.2	Optical filters for auto-stereoscopic displays: a) Lenticular sheet, b) Parallax barrier . . . . .	20
3.3	Optical filters for multiview auto-stereoscopic displays: a) Lenticular sheet , b) Parallax barrier . . . . .	21
4.1	Simultaneous 2D and 3D presentation of 3D content as introduced in [P2] . . . . .	32
4.2	2D presentation of stereoscopic video combinations from (a) original stereopair and (b) proposed rendered stereopair . . . . .	34
5.1	Asymmetric stereoscopic video . . . . .	35
5.2	Average subjective ratings and 95% confidence intervals for different eye dominant subjects . . . . .	36
5.3	Examples of different types of asymmetric stereoscopic video coding .	38
5.4	Encoding times for full and quarter resolution views . . . . .	40
5.5	Block diagram illustrating the placement of down and upsampling blocks for different applications . . . . .	41
5.6	Subjective test results for (a) low bitrate and (b) high bitrate sequences	45
5.7	Correlation between subjective scores and objective estimates . . . . .	51
6.1	Encoding and synthesis process for a depth-enhanced multiview video	54
6.2	A synthesized view . . . . .	57
6.3	Rendered view from (a) original depth map and (b) low-pass filtered depth map . . . . .	58
6.4	Resampled depth maps (a) original, (b) proposed method in [P1], (c) JSVM . . . . .	61
6.5	Encoding artifacts(a) blocking and (b) blurring . . . . .	64



# List of Tables

5.1	Spatial resolution of the sequences for different downsampling rates . . . . .	44
5.2	QP selection of different methods for the left view (right views are identical for different coding methods of each sequence) . . . . .	45
5.3	Tested bitrate values per view and the respective PSNR values achieved by symmetric stereoscopic video coding with H.264/AVC . . . . .	45
5.4	Statistical significance differences (SSD) of asymmetric methods against FR symmetric(1 = there is SSD, 0 = No SSD) . . . . .	46
5.5	Bitrate selection for different sequences . . . . .	49
5.6	Pearson correlation coefficient between VQM values and mean subjective scores . . . . .	50
6.1	PSNR of synthesized views based on spatial resolution of reference texture and depth views . . . . .	59





# Chapter 1

## Introduction

Currently a large quantity of video material is distributed over broadcast channels, digital networks, and personal media due to the ever increasing trend in video consumption. Such increase in popularity of the video content demands higher resolution and quality of the provided material. An obvious requirement for such a growing appetite is a more intelligent and efficient coding algorithms enabling the end users to access content with the highest subjective quality while respecting the limitations in the broadcasting and storage facilities. This is further complicated while changing the dimension of the video from conventional 2D to 3D, resulting in an increase in the number of pixels to be coded for the equivalent content to provide the subjects with depth perception of the scene similar to what is experienced in daily life. This is an inevitable trend in video content acquisition and creation since typically the user satisfaction increases while switching to 3D content from the traditional 2D content. The vast research and industrial activities on improving the 3D display technology, 3D acquisition, 3DV compression, and 3D movie making confirms the desire of the users in this regard. Since the evolution of content production, video acquisition/rendering, and display technologies is much faster than the networks and the broadcasting capabilities, an obvious requirement for the new video coding standard is identified. Such a new standard should target outperforming the current state-of-the-art H.264/AVC (the same as MPEG-4 Part 10) [117].

3D perception can be achieved by providing each eye with a slightly different view. These two views can be the reference views, i.e. the views which have been transmitted or can be output of some rendering algorithm applied to the reference views. In multiview video format several cameras capture the same scene from different points of view. Stereoscopic video is a subset of multiview format where only two of the views are utilized or generated. In the case of traditional stereoscopic video, MVC [29], as an annex to H.264/AVC, is the state-of-the-art and exploits inter-view redundancies while encoding different views. Several approaches are proposed to increase the efficiency of MVC e.g. harmonizing the views by removing the introduced noise during the capturing process [9], reducing the spatial resolution of

all or a subset of views targeting lower complexity and reduced required bitrate to encode the same content, or applying low-pass filter (LPF) to all or some of the views targeting less accuracy in high frequency components (while maintaining acceptable subjective quality) and hence, bitrate required for compression process [7].

Compared to conventional frame-compatible stereoscopic video coding as well as multiview video coding, depth-enhanced multiview video coding provides more flexibility in 3D displaying at the user side. While the availability of the two decoded texture views provides the basic 3D perception of traditional stereoscopic displays, it has been discovered that disparity adjustment between views is needed for adapting the content on different displays and for various viewing conditions, as well as bringing satisfaction to different individual presences [138]. Furthermore, since autostereoscopic display (ASD) typically requires a relatively large number of views simultaneously, it is not possible to transmit or broadcast such a huge amount of data under the current network capabilities. Therefore, the multiview video plus depth (MVD) format [141] is considered where each texture view is associated with a respective depth map, and only few depth-enhanced views are transmitted and the rest of the required views are rendered in the playback device using the depth image based rendering (DIBR) algorithms [86]. Depth-enhanced multiview video coding schemes can also benefit from possible approaches introduced for MVC as well as removing a subset of potential redundant depth views from the MVD package, as long as no significant drop in the subjective quality of the rendered views is introduced, targeting a bitrate reduction due to the smaller number of depth views to be encoded. Different formats used in this thesis to present 3D content are depicted in Figure 1.1.

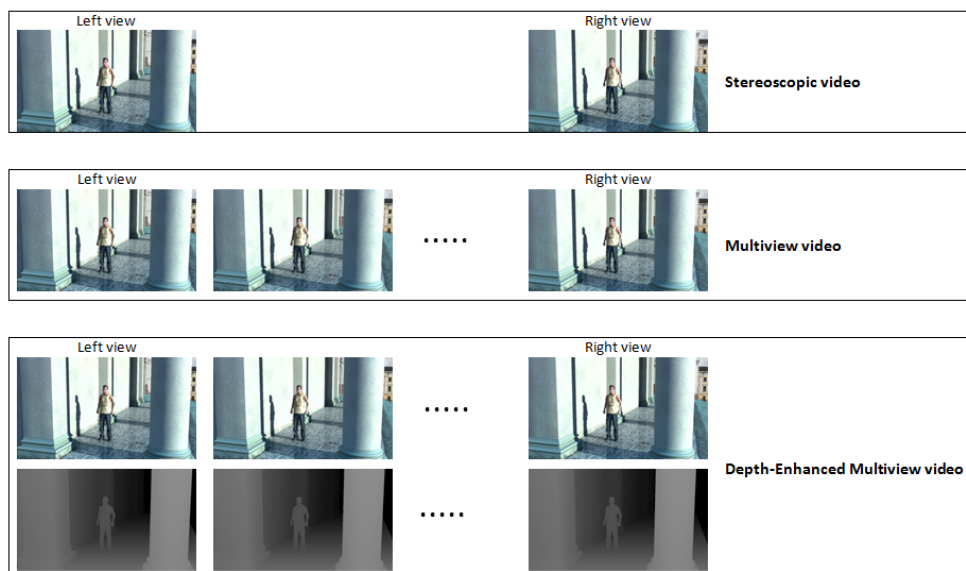


Figure 1.1: Different formats to present 3D content

---

One promising scheme to encode both stereoscopic and multiview content is to encode the views asymmetrically, i.e. the quality of all views is not degraded to the same extent and some views face more artifacts compared to other views. In this case, attributed to binocular suppression theory [15], the HVS is expected to fuse the perceived content in such a way that the higher quality view contributes more to the final observed subjective quality. However, despite abundant research and experiments, this concept is not still well comprehended and depends on several factors, e.g. the limits of asymmetry introduced to the views, the type of quality asymmetry, the viewing distance, and the degradation level applied to the views. Therefore, depending on the target applications and considering the content, the parameters tuning the asymmetry should be selected wisely to achieve the aimed performance.

All new coding proposals are conventionally compared to the state-of-the-art codec objectively, to reveal whether they provide a higher performance than the already available codec or not. Objective metrics are usually reliable and estimate accurately the subjective quality, however, they do not necessarily align with the HVS preference. This means, there might be a case where some content has a higher subjective quality while the objective metrics fail to estimate such a higher quality due to their potential limitations in estimating the HVS fusion process. For example, when a small spatial movement in the content grid happens or in the case where some high frequency components which are not subjectively visible are removed, non-perceptual objective metrics report a misleading estimation of subjective quality. Moreover, exploiting objective metrics ignores the conditions, the display, and the setup under which the content is perceived. Especially in the case of 3DV, where two views are provided, no objective metric is known to be able to precisely approximate the fusion process of our HVS and hence, it is obligatory to perform subjective quality assessment to assure a relatively more accurate evaluation of the proposed algorithms.

## 1.1 Objectives and outline of the thesis

This thesis focuses on various approaches in compression of different formats of 3D content and several potential techniques added to the reference codecs have been introduced/evaluated targeting a better efficiency compared to the reference codecs excluding the proposed techniques. A major contribution of the experiments and research presented in this thesis deals with the concept of asymmetry on video content where some of the views have a lower quality compared to the other view. A major objective of this thesis is to show that under different 3DV formats, targeting different types of displays, transmitting some views with coarser encoding techniques can provide users with a similar subjective quality compared to that offered by symmetric views. Obviously, this is achieved under some constraints on the level of asymmetry between views which is also discussed in this thesis.

The research presented in this thesis can be categorized into two categories. One category is to evaluate the proposed coding scheme on conventional stereoscopic video, containing only two views, targeting the highest subjective quality. This was achieved with different approaches including several asymmetric schemes. It was concluded that in general the evaluated asymmetric schemes present a promising approach to reduce the bitrate while maintaining the subjective quality of the corresponding symmetric video. The second category focuses on depth-enhanced multiview video targeting a higher objective and/or subjective quality for the stereopair created with coded and synthesized views. In this thesis, I am not targeting any view synthesis algorithm and always the state-of-the-art scheme is being used for both proposed and reference codecs. This includes novel algorithms for better compression of depth maps and new methods and schemes allowing more efficient encoding of texture views. Both categories in general deal with the compression of 3D content but in different formats and the stereoscopic video compression can be considered as a subset of the multiview video compression,

Some of the proposed schemes in this thesis have been evaluated objectively. However, since the concept of asymmetry in several studies has been utilized and the objective metrics were found unable to well estimate the perceived quality of asymmetric quality stereoscopic video [53], several subjective quality assessments were conducted in this thesis. The subjective evaluation results consistently confirmed that the proposed schemes outperform the analogous symmetric cases under the same bitrate constraint, or equivalently, they are able achieve similar subjective quality while decreasing the required bitrate. This is an important objective of this thesis to confirm a higher performance of the proposed encoding algorithms subjectively to guarantee accurate quality assessment.

The thesis is organized as follows. In chapter 2 HVS is described with a focus on the related concepts to this thesis. Following this brief overview, different types of displays, covering the targeted end-user devices related to the encoding methods proposed in this thesis, are introduced in chapter 3. In chapter 4, the quality evaluation of 3D content is explained by describing several objective metrics as well as subjective test criteria. Moreover, the subjective quality of 3D content displayed on traditional stereoscopic displays is analyzed when perceived with or without glasses. The concept of asymmetry in video compression is described in chapter 5. This chapter covers different types of asymmetry and justifies their utilization while discussing the criteria by which the level of asymmetry between views are limited. The conclusions based on the conducted subjective tests on asymmetric stereoscopic video are reported at the end of this chapter. In chapter 6, the depth-enhanced multiview video format is introduced to be used in DIBR algorithms for synthesizing views and it is explained how the quality of synthesized views varies based on the quality of the used texture and depth views. The compression of depth-enhanced multiview format is further discussed in this chapter with an emphasis on having asymmetric quality between the views. Moreover, the subjectively confirmed conclusions regarding this 3D content format are presented at the end of this chapter.

---

Finally, conclusions and future works are drawn in chapter 7.

## 1.2 Publications and author's contribution

This thesis is based on the publications that represent original work in which the thesis author has been the essential contributor. Considering that all publications included in the thesis are the outcome of team work, the author's contribution to each publication is described in the following paragraphs. All publications are written mainly by the thesis author while reviews, comments, and modifications are provided by co-authors. Moreover, all simulations required for publications are performed by thesis author except for [P4].

In [P1], a novel non-linear method, co-invented by thesis author, Miska Hannuksela, and Dmytro Rusanovskyy, for depth map resampling is introduced. Thesis author has implemented the idea and written the paper.

[P2] proposes a novel technique to present the content of 3D displays so that subjects with and without glasses are able to simultaneously perceive high quality 3D and 2D content, respectively. Such proposal has not been introduced to the research community before and is considered to have a potential bright future for researchers working in this field. Thesis author co-invented the idea with Miska Hannuksela and the algorithm was implemented by the thesis author. A software was implemented by Hamed Sarbolandi to conduct the subjective tests. Thesis author has analyzed the subjective scores and written the paper.

We gathered a summary of previous publications written by thesis author on subjective quality assessment of asymmetric stereoscopic video in [P3] by introducing a more comprehensive deepened analysis of the statistics and results. A set of conclusions are drawn in this article and hence, it is considered to be a proper reference for future subjective quality evaluation research concerning asymmetric quality in stereoscopic video compression. Thesis author has written the paper.

A new MVD format to represent the multiview plus depth 3D content is introduced in [P4] and thesis author has performed the required modifications to infrastructure to enable the support for the proposed format. Paper is written by thesis author.

In [P5], a new asymmetric scheme for multiview video content is proposed by authors and changes in the test software to support such scheme were implemented by thesis author and Wenyi Su. Thesis author has written the paper.

Targeting a new MR asymmetric scheme, thesis author, Miska Hannuksela, and Moncef Gabbouj co-invented a format which is introduced in [P6]. The subjective evaluation compares the quality of this format with conventional MR scheme and FR stereoscopic video. Subjective assessment is conducted by Maryam Homayouni while rating analysis and writing the paper was done by thesis author.

Considering the amount of high frequency components in the texture views, a new method is presented in [P7], aiming to decide which spatial resolution enables

---

a more efficient encoding for multiview 3D content. Thesis author has proposed the algorithm and implemented it. Subjective tests are performed in Human-Centered Technology of Tampere University of Technology and the thesis author has written the paper.

We propose a scheme consisting of asymmetric quality among different views in a depth-enhanced video in [P8] and considering lower quality of some views, lower bitrate compared to anchor, where all views have full-resolution (FR), is achieved. The subjective quality assessments were conducted in Human-Centered Technology of Tampere University of Technology and the thesis author has written the paper.

In [P9] a new mixed-resolution (MR) scheme is proposed where sample value quantization and spatial resolution adjustment are used together to create asymmetry between views of stereoscopic video targeting better compression. Miska Hannuksela and thesis author have proposed the algorithm and thesis author has implemented it. The subjective tests were conducted by department of media technology in Aalto univeristy and the paper was written by thesis author.

A new model to estimate the subjective quality of MR stereoscopic video is proposed by thesis author in [P10] and he has evaluated the efficiency of the proposed metric taking into account the results of two sets of subjective tests under different test setups. The subjective tests were performed by department of media technology in Aalto University and the thesis author has written the paper.

## Chapter 2

# Human Visual System

The HVS consists of several organs, e.g. the eyes, the nerves, and the brain. The whole concept of the HVS can be discussed from two different points of view, the visual perception and visual cognition. Visual perception is a subject of anatomy [62, 167] while visual cognition as a higher level processing function of the brain is studied in psychology [26, 167].

The functioning of a camera is often compared with the workings of the eye; both focus light from external objects in the visual field onto a light-sensitive screen. Analogously to a camera that sends a message to produce a film, the lens in the eye refracts the incoming light onto the retina. Several optical and neural transformations are required to provide visual perception. The retina is made up by millions of specialized photoreceptors known as rods and cones. Rods are responsible for vision at low light levels (scotopic vision). They do not mediate color vision and have a low spatial acuity and hence, are generally ignored in the HVS modeling [167]. Cones are active at higher light levels (photopic vision). They are capable of color vision and are responsible for high spatial acuity. There are three types of cones which are generally categorized to the short-, middle-, and long-wavelength sensitive cones i.e. S-cones, M-cones, and L-cones, respectively. These can be thought by an approximation to be sensitive to blue, green, and red color components of the perceived light. Each photoreceptor reacts to a wide range of spectral frequencies, with the peak sensitivity at approximately 440nm (blue) for S-cones, 550nm (green) for M-cones, and 580nm (red) for L-cones. The brain has the ability to fetch up the whole color spectrum from these three color components. This theory known as trichromaticism [153] allows one to construct a full-color display using only a set of three components. Despite the fact that perception in typical daytime light level is dominated by cone photoreceptors, the total number of rods in the human retina (91 million [102]) far exceeds the number of cones (roughly 4.5 million [102]). Hence, the density of rods is much greater than cones throughout most of the retina. However, this ratio changes dramatically in the fovea placed in the center of the projected image which is the highly specialized region of the retina measuring about 1.2 millimeters in diameter. The increased density of cones in the fovea is accompanied by



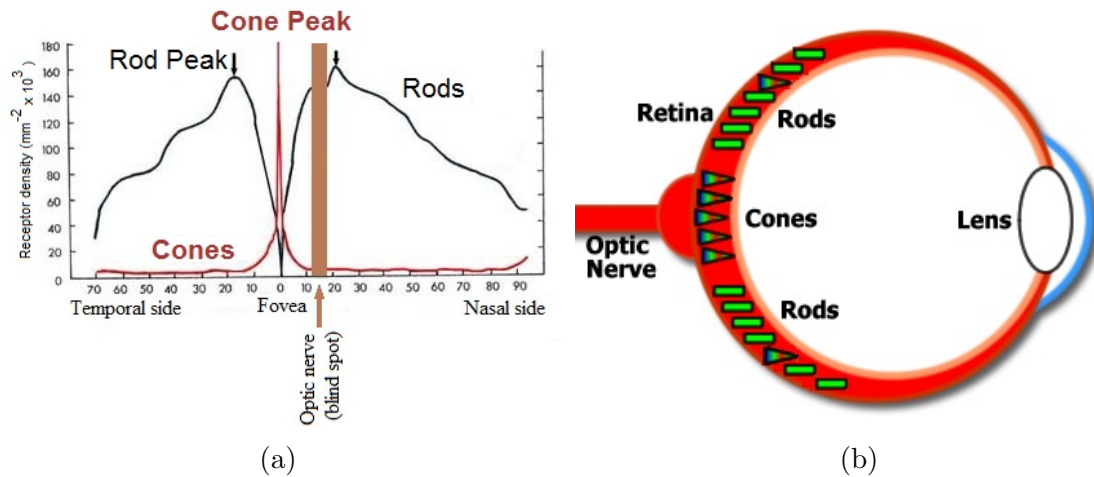


Figure 2.1: Cones and rods in the retina

a sharp decline in the density of rods. This is depicted in Figure 2.1. For further information regarding the structure of the retina readers are referred to [64].

The optic nerves leave the eye in a special region of the retina commonly known as the *blind spot* where no photoreceptors are available. As a result, there is no response to the light stimulus at this point and hence, the brain gets no information from the eye about this particular part of the projected picture. Light entering the eye is refracted as it passes through the cornea and the amount of light is adjusted by the pupil (controlled by the iris). This optical system of the eye in collaboration with a sensitivity adaptation mechanism in the retinal cells enables the eye to work over a wide range of the luminance values. In general, the eye is sensitive only to the relative luminance change (i.e. contrast), rather than absolute luminance values [87].

Light strikes the rod and cone cells causing electrical impulses to be transduced and transmitted to the bipolar cells. The processing in the retina includes the formation of bipolar and ganglion cells in the retina, as well as the convergence and divergence from photoreceptor to the bipolar cell. In addition, other neurons in the retina, particularly horizontal and amacrine cells, transmit information laterally (from a neuron in one layer to an adjacent neuron in the same layer), resulting in more complex respective fields that can be either indifferent to color and sensitive to motion or sensitive to color and indifferent to motion. The reticular activating system and bipolar cells in turn transmit electrical activity to the central nervous system from blind spot (where long ganglion cell axons exit the eye) and through the optic nerve [64] (see Figure 2.1). Each eye has about one million fibers [47]. Most of the fibers of the optic nerve terminate in the lateral geniculate nucleus (LGN) from where information is relayed to the visual cortex.

There are two main types of cells in the LGN: the first set of the cells are substantially larger than the other type of cells and are called *magno cells*. The main

inputs to these cells are the retinal rods and the magno ganglion cells. The cells in the magnocellular layers seem to be mainly responsible for transmitting information about motion and flicker perception, stereopsis, and high contrast targets (high temporal and low spatial resolution). The other type includes cells which are smaller and are called *parvo cells*. The main input to these cells is the retinal cones and the parvo ganglion cells. These cells are mainly responsible for transmitting information about the visual acuity, the form vision, the color perception, and the low contrast targets (slow response but high resolution in space). Such separation of cell types allows LGN to encode the motion information using a temporal resolution of as little as 10 to 12 frames per second [113].

## 2.1 Binocular human vision

Binocular vision is the ability to perceive visual information through two eyes. Human eyes are separated horizontally by a distance of approximately 6.3 cm on average [62]. Such positioning enables each eye to see the world from a slightly different perspective (Figure 2.2). There are 6 muscles that control the movement of the eye [23]. Four of the muscles control the movement in the cardinal directions i.e. up, down, left, and right. The remaining two muscles control the adjustments involved in counteracting head movement. To maintain single binocular vision when viewing an object, a simultaneous movement of both eyes toward each other is needed to enable convergence. Tracking describes the ability of the eyes to converge and hold on to an object even when the object is moving.



(a) Left View

(b) Right View

Figure 2.2: Left and right perspective of stereoscopic content

The HVS perceives color images using receptors on the retina of the eye which respond to three broad color bands in the regions of red, green and blue (RGB) in the color spectrum as explained in the previous section. The HVS is much more sensitive to the overall luminance changes than to color changes. The major challenge in understanding and modeling visual perception is that what people see is not simply a translation of the retinal stimuli (i.e., the image on the retina). Moreover,

the HVS has a limited sensitivity; it does not react to small stimuli, it is not able to discriminate between signals with an infinite precision, and it also presents saturation effects. In general one could say that the HVS achieves a compression process in order to keep the visual stimuli for the brain within an interpretable range. While presenting different views for each eye (stereoscopic presentation), the subjective result is usually *binocular rivalry* where the two monocular patterns are perceived alternately [174]. In particular cases, one of the two stimuli dominates the field. This effect is known as *binocular suppression* [74, 165]. It is assumed according to the binocular suppression theory that the HVS fuses the two images such that the perceived quality is close to that of the higher quality view at any time.

Binocular rivalry affords a unique opportunity to discover aspects of perceptual processing that transpires outside of the visual awareness. In a stereoscopic presentation, the brain registers slight perspective differences between the left and right views to create a stable, 3D representation incorporating both views. In other words, the visual cortex receives information from each eye and combines this information to form a single stereoscopic image. Left- and right-eye image differences along any one of a wide range of stimulus dimensions are sufficient to instigate binocular rivalry. These differences include changes and variations in color, luminance, contrast polarity, form, spatial resolution, or velocity. Rivalry can be triggered by very simple stimulus differences or by differences between complex images. Stronger, high-contrast stimuli lead to stronger perceptual competition. Rivalry can even occur under dim viewing conditions, when light levels are so low that they can only be detected by the rod photoreceptors of the retina. Under some conditions, rivalry can be triggered by physically identical stimuli that differ in appearance owing to simultaneous luminance or color contrast. Therefore, the problem of how an image may be perceived when it is viewed with both eyes as a stereoscopic image is not fully understood yet. If both views are provided with equal quality, the perceived quality of the stereoscopic image is proportional to the quality of both views. On the other hand, if the quality or other factors of the left and right views differ, the HVS plays the main rule on defining the perceived quality of the stereoscopic image and dominating it with more details from a selected respective view.

## 2.2 Spatial perceptual information

Different contents are subject to different spatial complexities. The ITU-T Recommendation P.910 [114] proposes the metric Spatial Information (SI) to measure the spatial perceptual detail of a picture (2.1). The value of this metric usually increases for more spatially complex scenes. Based on this recommendation and utilization of the Sobel filter (2.2), SI along the vertical or horizontal direction can be measured separately. SI includes the quantity and the strength of the edges in different directions.

$$SI = \max_{time} \{std_{space}[Sobel(F_n)]\} \quad (2.1)$$

$$H_{Sobel} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (2.2)$$

The functional model of the binocular vision is shown in Figure 2.3. When the eye is relaxed and the interior lens is the least rounded, the lens has its maximum focal length for distant viewing. As the muscle tension around the ring of muscle is increased and the supporting fibers are thereby loosened, the interior lens rounds out to its minimum focal length. This enables the eye to focus on objects at various distances. This process is known as *accommodation* [158], and the refractive power is measured in diopters. Accommodation can be defined as the alteration of the lens to focus the area of interest on the fovea, a process that is primarily driven by blur [148,150]. Vergence deals with obtaining and maintaining a single binocular vision by moving both eyes, mainly in opposite directions. Naturally, accommodation and vergence systems are reflexively linked [21,108,123,127]. The amount of accommodation required to focus on an object, changes proportionally with the amount of vergence needed to fixate that same object in the center of the eyes. The cornea provides two third of the refractive power of the eye and the rest is provided by the

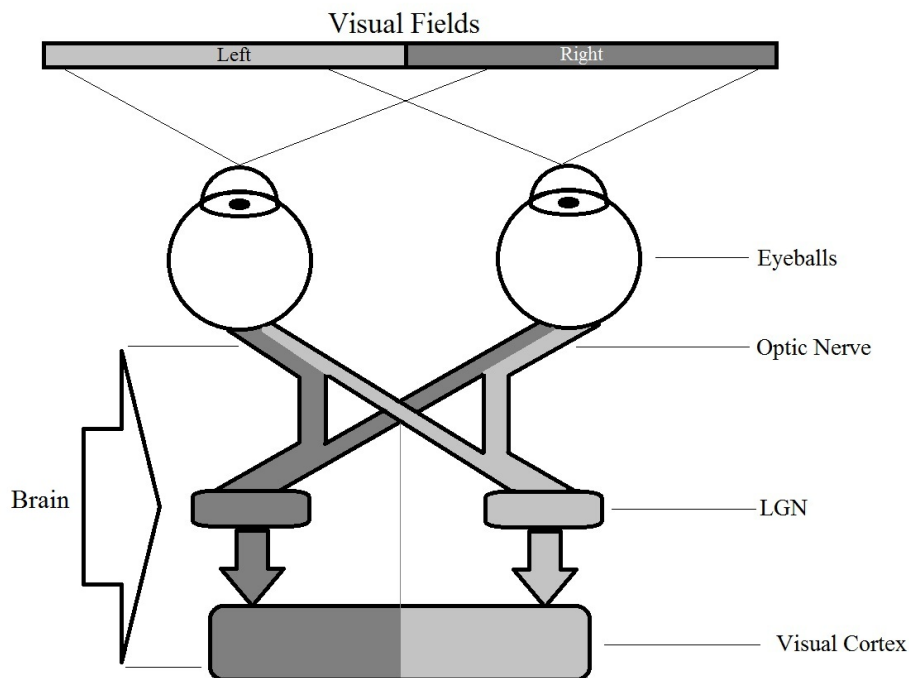


Figure 2.3: Functional model of binocular vision

lens. However, our eye tends to change the curvature of the lens rather than that of the cornea. Normally, when our ciliary muscles are relaxed, parallel rays from distant objects will converge onto the retina. If our eye is maintained at the above state, and a near object is put before it, light rays will converge behind the retina. As the sharp image is behind the retina, our brain can only detect a blurry image. To bring the image into focus, the eye performs accommodation. In cases where the optical system is unable to provide a sharp projected image, the blurring artifact is modeled as a low-pass filter characterized by a point spread function (PSF) [179]. When focusing near an object, the ciliary muscle contracts, and suspends the eye. As a result, surfaces of the cornea and the lens become more curved and thus the eye focuses on the nearby object. When two different perspectives of the scene are available in retinas of each eye, we call this *binocular disparity* [62]. The HVS utilizes binocular disparity to deduce information about the relative depth between different objects. The capability of the HVS to calculate depth for different objects of each scene is known as *stereovision*. For a certain amount of accommodation and vergence, there is a small range of distances at which an object is perfectly focused and a deviation in either direction gradually introduces blur. An area defining an absolute limit for disparities that can be fused in the HVS is known as *Panum's fusional area* [32, 112]. It describes an area, within which different points projected on the left and right retina produce binocular fusion and sensation of depth. Panum's fusional areas are basically elliptical having their long axes located in horizontal direction [91]. This is depicted in Figure 2.4.

The limits of Panum's fusional area are not constant over the retina, but expand

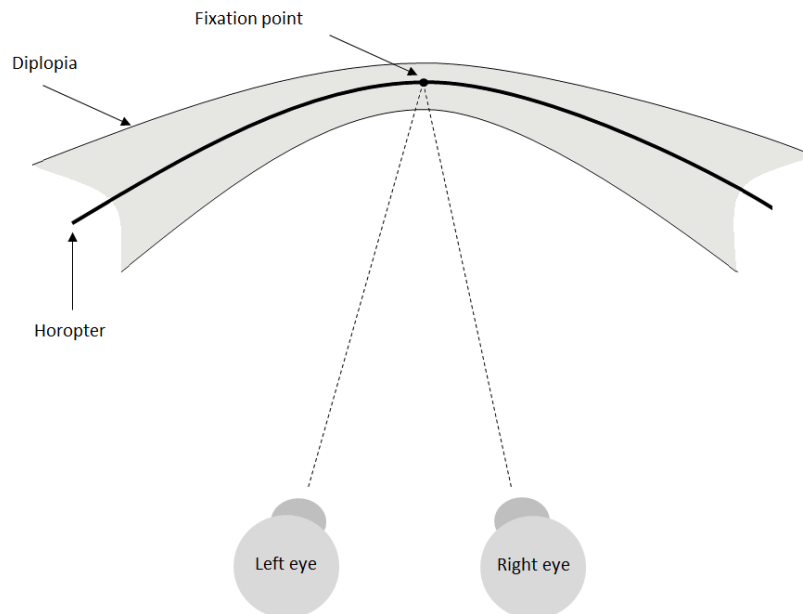


Figure 2.4: Panum's fusional areas

while increasing the eccentricity from the fovea. The limit of fusion in the fovea is equal to the maximum disparity of only one-tenth of a degree, whereas at an eccentricity of 6 degrees, the maximum value is limited to one-third of a degree [61, 173] and at 12 degrees of eccentricity without eye movements the maximum disparity is about two-third of a degree [104].

Considering the amount of light entering the eye and the sensitivity adaptation of the retina, our eye is able to work over a wide range of intensities between  $10^{-6}$  and  $10^{18}$  cd/m<sup>2</sup>. The fact that the eye is sensitive to a luminance change (i.e. contrast) rather than the absolute luminance is known as *light adaptation* and is modeled by a local contrast normalization [171]. The light projected onto the fovea that comes from the visual fixation point and has the highest spatial resolution is called *foveal vision*. The resolution of the surrounding vision to the foveal vision decreases rapidly and is known as the peripheral vision. Usually a non-regular grid is used to resample the image in a process known as *foveation* [73]. Due to different algorithms with which the visual information is processed, the HVS has a different sensitivity to patterns with different densities. The minimum contrast that can reveal a change in the intensity is called a *threshold contrast* and depends on the pattern density with a contrast sensitivity function (CSF) [167, 179]. The neurons in the visual cortex are sensitive to particular combinations of the spatial and temporal frequencies, spatial orientation, and directions of motion. This is well-approximated by two dimensional Gabor functions [167, 179]. To perceptually optimize the compression of images, the spatially dependent CSF is used [2].

The LGN receives information directly from the ascending retinal ganglion cells via the optic tract and from the reticular activating system. Both the LGN in the right hemisphere and the LGN in the left hemisphere receive input from each eye. However, each LGN only receives information from one half of the visual field, as illustrated in Figure 2.3. This occurs due to axons of the ganglion cells from the inner halves of the retina (the nasal sides) decussating (crossing to the other side of the brain) through the optic chiasm. The axons of the ganglion cells from the outer half of the retina (the temporal sides) remain on the same side of the brain. Therefore, the right hemisphere receives visual information from the left visual field, and the left hemisphere receives visual information from the right visual field. This information is further processed inside LGN.

The number of visual nerves going out of the LGN is about 1% of the neurons entering LGN. This suggests that in LGN a huge de-correlation of the visual information is performed including binocular masking and extraction of binocular depth cues. LGN fuses two input views to one output view called *cyclopean image* representing the scene from a point between the eyes. This image is then carried by the LGN axons fanning out through the deep white matter of the brain as the optic radiations, which will ultimately travel to the primary visual cortex (V1), located at the back of the brain. The binocular suppression theory and also anatomical evidence suggest that a small part of the visual information received in each eye might be delivered to V1 without being processed in LGN.

## 2.3 Binocular suppression theory

This section deepens the concept introduced in section 2.1 and further describes the conditions under which binocular suppression happens.

Binocular fusion occurs when a single binocular percept is produced by similar lights striking corresponding parts of each retina. The mechanism of the underlying fusion is imperfectly understood. One held interpretation of fusion assumes that the monocular inputs contribute equally to the production of an emergent single percept. Another alternative interpretation is the binocular suppression theory asserting that fusion results from the suppression or inhibitory interaction of the monocular images.

The binocular rivalry as a type of perceptual processing is resolved early in the visual pathway, resulting from mutual inhibition between monocular neurons in the primary visual cortex [16]. The perceptual dominance is influenced by the strength of each stimulus i.e. the amount of motion or contrast in each view. This is sometimes termed Levelt's 2nd proposition [16, 75]. Moreover, an addition of a contextual background can increase the predominance of the inconsistent target. Multiple stages of mutual inhibition between neural populations happen in the HVS. The neurons generating the dominant image inhibit the neurons corresponding to the suppressed image, but over time the system fatigues and the strength of inhibition reduces allowing the suppressed image to become dominant. This process continues indefinitely [16, 177].

In normal vision, there is some additional fusion to impulses from corresponding points of the two retinas. The correspondence of the retinal elements is completely rigid and un-changing; however, one of a pair of the corresponding points always suppresses the other. In the presence of a contour, the suppressing power of retinal elements on its sides is enhanced. In places where there is disparity of the contour in one eye, then the eye retinal elements on both sides of this contour will suppress corresponding points in the other eye. *Diplopia* happens when the extent of the suppression is smaller than the disparity between the contours, but still depth perception is expected. If the extent of the suppression is greater than the disparity between the contours, one contour is suppressed and single vision occurs with depth perception. It is possible that the contour of one part of the image may be dominant in one eye, and that of another part may be dominant in the other eye. According to the suppression theory, one of a pair of corresponding points always suppresses the other, and it would consequently be anticipated that binocular mixtures of colors could not occur. This is attributed to the widely believed assumption of the binocular suppression theory [15], which claims that the stereoscopic vision in the HVS fuses the images of a stereopair so that the visual perceived quality is closer to that of the higher quality view.

Several subjective quality evaluation studies have been conducted to research the utilization of the binocular suppression theory in asymmetric quality stereoscopic video [11, 20, 105, 142, 152]. We shall return to this topic in more details in chapter 5.

# Chapter 3

## 3D Content Visualization

This chapter provides information about scene characteristic and introduces different 3D displays describing how they are used for different required scenarios. A variety of display devices providing 3D experience have been commercialized. Among the 3D display solutions are stereoscopic displays requiring the use of polarizing or shutter glasses, and multiview ASDs, where the views seen depend on the position of the viewer relative to the display without requirement of viewing glasses.

### 3.1 Scene characteristics

Each scene can be characterized from several different perspectives. One point of view is the 3D visualization, describing the content with different depth sensations compared to the position of the viewer. This is one of the most familiar concepts for scene visual assessment and is experienced daily by all of us. The result is to see what happens around us knowing that e.g. how close is some particular object to us and whether it is moving toward or from us. Recently, considering the improvements in 3D visualization, many companies and research centers are actively involved in 3D video exploiting especially the need of users to watch movies, play games, and communicate with devices in 3D. This is due to the fact that these devices provide analogous feeling to users as if they were actually in the location of the scene, since a similar depth perception feeling is created.

There has been some effort on providing an efficient technique to enhance 3D videos by reducing the feeling of artificial clarity (including motion and disparity information of 3D contents) which can be experienced by the viewers [186]. The authors in [96] accomplish such aim by taking into account some characteristics of the human visual perception to define a joint motion-disparity processing approach, which is employed to enhance 3DV contents by reducing the feeling of artificial clarity, and thus resulting in an improved user acceptance and satisfaction. In the following sections different 3D displays are introduced and briefly explained.



## 3.2 3D displays

An important first step towards a high quality 3D display system is defining the requirements for its hardware and the images shown on it. Binocular vision provides humans with the advantage of depth perception derived from the small differences in the location of the similar points of the scene on the retina of the left and right eyes. Precise information of the depth relationships of the objects in the scene are provided by stereopsis. The HVS also utilizes other depth cues to help interpret the two images. These include monocular depth cues, also known as pictorial [51] and empirical [98] cues, whose significance is learnt over time, in addition to the stereoscopic cue [98].

People with monocular vision are able to perform well when judging depth in the real world. Therefore, 3D displays should be aware of the major contribution of monocular 2D depth cues to depth perception and aim to provide at least as good a basic visual performance as 2D displays. In [45] it is suggested that this should include levels of contrast, brightness, resolution, and viewing range that match a standard 2D display with the addition of the stereoscopic cue providing depth sensation through a separate image for each eye.

Wheatstone in 1838 [174] demonstrated that the stereoscopic depth feeling could be recreated by showing each eye a separate 2D image. Wheatstone was able to confirm this feeling by building the first stereoscope and many devices have been invented since then for stereoscopic image presentation having their own optical configurations. Reviews of these devices and the history of stereoscopic imaging are available in several sources, [13, 58, 72, 80, 161].

## 3.3 Stereoscopic displays

Stereoscopic displays require users to wear a device, such as analyzing glasses, to ensure that left and right views are seen by the correct eye. Many stereoscopic display designs have been proposed and there are reviews of these in numerous reports [13, 58, 80, 84, 161]. Most of these are mature systems and have already become established in several markets, as stereoscopic displays are particularly suited to multiple observer applications such as cinema and group presentation. Hence, it seems that the display solutions based on glasses are more mature for mass markets and many such products are entering the market currently or soon.

The lenses of polarizing glasses used for stereoscopic viewing have orthogonal polarity with respect to each other. The polarization of the emitted light corresponding to pixels in the display is interleaved. For example, odd pixel rows might be of a particular polarity, while even pixel rows are then of the orthogonal polarity. Thus, each eye sees different pixels and hence perceives different pictures. The shutter glasses are based on active synchronized alternate-frame sequencing. There is a synchronization signal emitted by the display and received by the glasses. The synchronization signal controls which eye gets to see the picture on the display and

for which eye the active lens blocks the eye sight. The left and right view pictures are alternated in such a rapid pace that the HVS perceives the stimulus as a continuous stereoscopic picture and therefore, depth sensation is provided.

### 3.3.1 Passive displays

Passive 3D displays require glasses with special lenses that filter images associated to each eye to produce a 3D sensation. The two pictures are shown superimposed on each other, with a filter on the screen to make the two pictures distinct. Watching such a display, the filters in the glasses guarantee that each eye only sees the respective image that it is supposed to see. Viewing glasses are classified to different categories based on the type of filters used. One solution is to exploit different filters with usually chromatically opposite colors. This is known as *anaglyph 3D glasses* and when the filtered content passes through the glasses, an integrated stereoscopic image is revealed to the HVS. Another more popular type of glasses is polarizing glasses where the glasses contain a pair of different polarizing filters. Each filter only passes the light that has been similarly polarized and blocks the light polarized in the opposite direction. Either orthogonal or circular polarizing filters for separating the left and right eye view are utilized in polarized glasses.

Polarized glasses have the advantage that full color and refresh rate is perceived, but the disadvantage is that special display hardware is required. In row interlaced polarized displays, every other row is presenting the content of the left or right view. Hence, since the vertical spatial resolution of the polarized display should be divided between the left and right views, the perceived spatial resolution of each view is half of the actual vertical resolution. Therefore, depending on the content, display technology, and the software playing the 3D content, if a proper low-pass filtering is not applied prior to the presentation of each view with half vertical resolution of the display, an annoying aliasing artifact [33] might be visible.

Passive displays are more independent compared to active displays and do not require any output device to synchronize their refresh rate. Passive displays require polarized glasses which do not have any electronics or power needs, and therefore, they are very light and inexpensive; but initial cost of the display itself is often greater than the equivalent active 3D display. Moreover, as long as the cost is the main factor, the passive method of displaying stereoscopic images is better suited for large groups, since the expensive technology is primarily in the display rather than in the glasses.

### 3.3.2 Active displays

Active 3D displays require glasses with electronic shutters that flicker in time, synchronized with the frequency of the display, to separate the picture into two images (or frames). The screen rapidly shows the left and right pictures, and a built-in infrared emitter or radio transmitter tells the glasses how fast they have to shutter

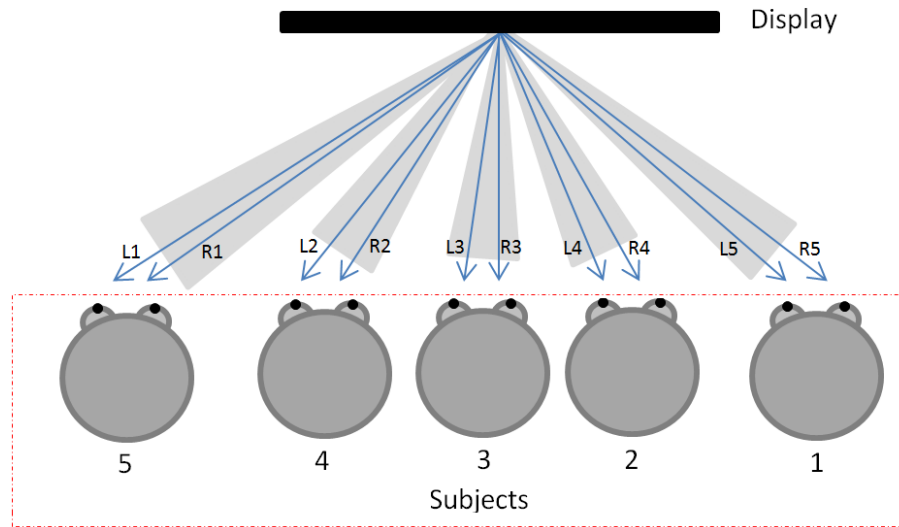


Figure 3.1: Auto-stereoscopic display

to make sure each respective image is delivered only to the corresponding eye. Each image is only visible to a one eye, giving the effect of depth to the viewer.

The glasses are electronic devices including a receiver and power supply, so they tend to be bulkier, less comfortable, and more expensive compared to passive glasses. They mostly eliminate the cross-talk [71] which might be present in passive displays, and as a result the same content is expected to have a higher subjective quality and 3D perception in active displays compared to passive ones. However, active glasses have the advantage that the 3D content is perceived with the FR and color, but the disadvantage is the necessity of active glasses and displays with very high refresh rates to guarantee nonexistence of flicker. If, for instance, the display supports frequency of 120 Hz, each view will have a refresh rate of 60 Hz.

### 3.4 Auto-stereoscopic displays

ASDs offer the viewer 3D realism close to what is experienced in the real world. In real life we gain 3D information from a variety of cues. Two important cues are stereo parallax i.e. seeing a different image with each eye, and movement parallax i.e. seeing different images when we move our heads. ASDs combine the effects of both stereo and movement parallax in order to produce the perceived effect similar to that of a white light hologram [37]. Figure 3.1 shows the viewing space in front of an ASD divided into a finite number of horizontal zones. In each zone only one stereo pair of the scene is visible. However, each eye sees a different image and the images change when the viewer moves his head between zones.

ASDs are a class of 3D displays which create depth effect without requiring the observer to wear special glasses. Such displays use additional aligned optical elements on the surface of the screen, ensuring that the different images are delivered

to each eye of the observer. Typically, ASDs can present multiple views to the viewer, each one seen from a particular viewing angle along the horizontal direction, creating a comfortable viewing zone in front of the display for each pair of views. However, the number of views comes at the expense of resolution and brightness loss. One key element that influences the perceived performance of ASDs is the subjective quality of the viewing windows that can be produced at the nominal viewing position. The quality of respective viewing windows can degrade due to unresolved issues in the optical design leading to flickering in the image, reduced viewing freedom, and increased inter-channel cross-talk. These can reduce the quality of viewing experience for observers in comparison to the stereoscopic 3D displays.

In general, due to the use of glasses, the 3D perception quality in ASDs is lower compared to stereoscopic displays. Considering the number of views provided by ASDs, they are categorized in two different classes, as explained in the following sub-sections.

### 3.4.1 Dual-view auto-stereoscopic displays

In Dual-view ASD, two images are transmitted and each is visible from a different perspective. There exist several observation angles and if correctly positioned, the observers are able to see the 3D content from different viewing zones. Figure 3.1 shows a typical dual-view ASD where a finite number of zones in which a stereopair is perceived, are created in front of the display.

To enable one display beaming two different images, several approaches have been proposed of which the most common approach is to put an additional layer in front of the thin film transistor liquid crystal display (TFT-LCD) [66,103,147]. This layer alters the visibility of display sub-pixels, and makes only half of them visible from a given direction. This layer, known as optical filter [159] has two common types: lenticular sheet [162] and parallax barrier [159]. Lenticular sheet is an array of magnifying lenses, designed to refract the light to different directions as shown in Figure 3.2a [163]. Parallax barrier consists of a fine vertical grating placed in front of a specially designed image, so it is basically blocking the light in certain directions as shown in Figure 3.2b [66]. In both optical filter types, considering that only half of the available sub-pixels on display are perceived with each eye, the resolution of the perceived view by each eye is lower than the 2D resolution of the display.

### 3.4.2 Multiview auto-stereoscopic displays

Multiview ASDs typically work in a similar way to the spatially-multiplexed dual-view ASDs. However, instead of dividing the sub-pixels to only two views, typically 8 to 28 views are created. As for light distribution techniques, the same lenticular sheets [162] or parallax barrier [159] are utilized. Lenticular sheet refracts the light while parallax barrier blocks the light in certain directions, as shown in Figures 3.2a

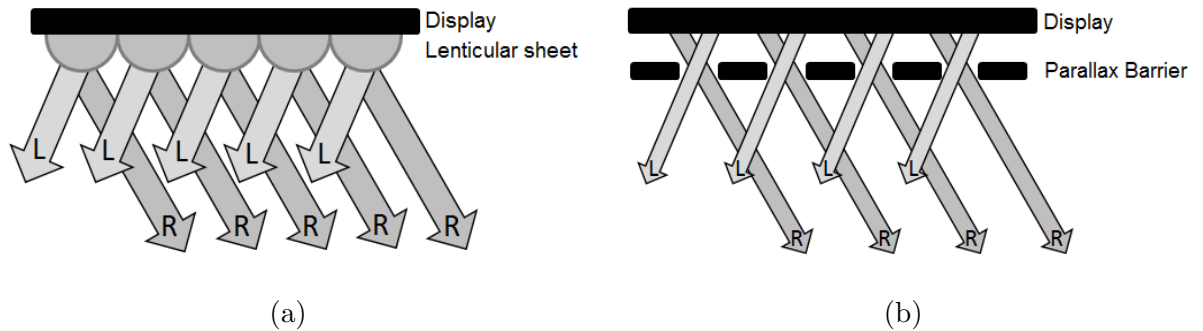


Figure 3.2: Optical filters for auto-stereoscopic displays: a) Lenticular sheet, b) Parallax barrier

and 3.2b, respectively.

Applying the optical filter limits the maximum perceived brightness of each sub-pixel to a certain angle called *optimal observation angle* for that sub-pixel. The optical observation angles of different sub-pixels for the same view are designed to intersect in a narrow spot in front of the display. This spot tends to have the highest brightness for that view. Moving sideways from this spot, still the view is visible with a diminished brightness. The window in which the view is still visible is called *visibility zone* of the view and in most multiview displays the visibility zones are located horizontally in front of the display. In the horizontal structure, visibility zones appear in fan shaped configuration similar to what is shown in Figure 3.1. The last view of each visibility zone is followed by the first view of the adjacent visibility zone. Hence, one central set of visibility zones are created directly in front of the display and a number of identical sets are repeated.

Considering that the number of pixels available in the display is limited, there exists a trade-off between the resolution of each view and the number of views provided by the display. Since generally depth cues are perceived in the horizontal direction, many multiview display producers do not allocate pixels for extra vertical views [39,147,159,162]. The advantages of such an approach is that the viewers are free to place their head anywhere within the visibility zone, while still perceiving a 3D image. Also, the viewer can “look around” objects in the scene simply by moving the head. Moreover, multiple viewers can be supported, each seeing 3D from a desired own point of view (see Figure 3.3), discarding the requirement to head-tracking with all its associated complexity. However, there are a few disadvantages for multiview ASDs, from which we can mention the difficulty of building a display with many views and also the problem of generating all the views simultaneously [25], because each view is always being displayed regardless whether or not it is seen by anyone. The behavior of an ideal multiview ASD is completely determined by four parameters: the screen width, the visibility zone width, the number of views, and the optimal viewing distance [38]. Considering the glasses-free approach used in

ASDs and further improvements introduced in multiview ASDs providing the users with more freedom to select an appropriate viewing point in front of the display, multiview ASDs tend to be a potentially promising choice for future 3D displays.

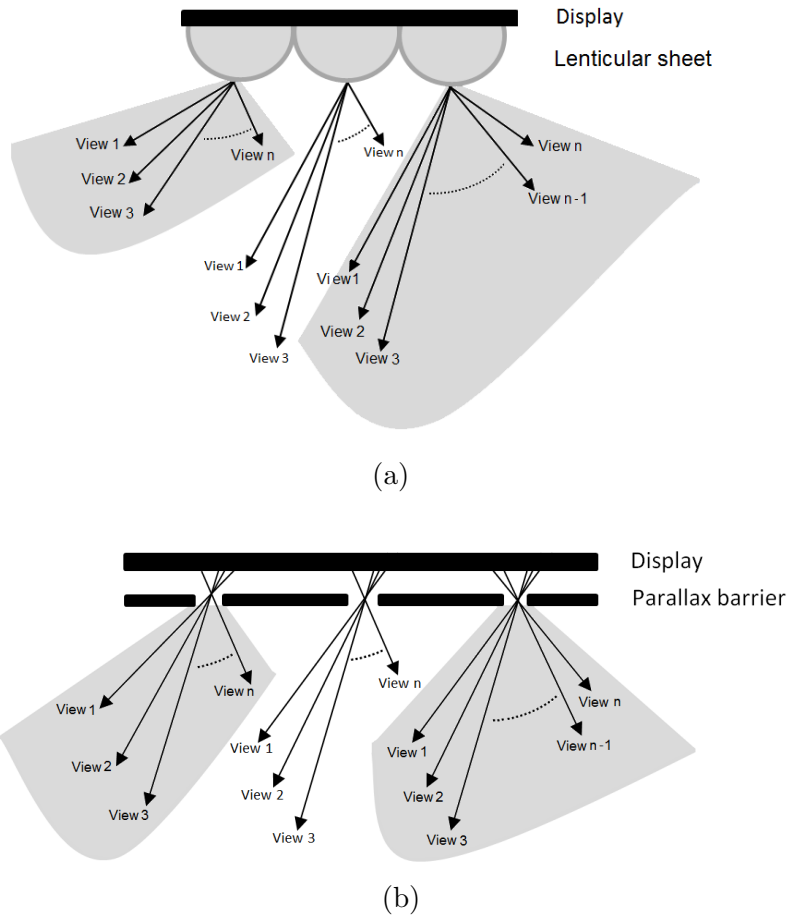


Figure 3.3: Optical filters for multiview auto-stereoscopic displays: a) Lenticular sheet , b) Parallax barrier



# Chapter 4

## Quality Assessment of 3D Video

Digital images typically undergo a wide variety of distortions from acquisition to transmission and display, which usually result in the degradation of the subjective quality. Hence, image quality assessment (IQA) is an essential approach to calculate the extent of the quality loss. Moreover, IQA is used to evaluate the performance of processing systems e.g. different codecs and enables the selection of different tools and their associated parameters to optimize the processing steps. There has been extensive research introducing new objective metrics [31, 79, 181] to evaluate the subjective quality of images.

For the majority of processed digital images, the HVS is the ultimate receiver and is the most reliable way of performing the IQA and evaluate their quality based on subjective experiments (defined in ITU-R Recommendation BT.500 [115]). Subjective evaluation is in general time consuming, expensive, and cannot be repeated. Hence, the usage of subjective evaluation is limited and cannot be conducted for the majority of the assessment scenarios. However, subjective quality assessment is still the most trustable approach to evaluate different processing algorithms and, for the cases where objective metrics fail to accurately estimate the visual quality or there is need of more precise evaluations, it remains the only choice. Yet, the existence of the limitations mentioned above has triggered a trend to develop objective IQA measures that can be easily embedded in the current systems. Some of these objective metrics are introduced and discussed in the next section.

While objective metrics are unable to accurately estimate the subjective quality of single-view images, this problem is boosted when stereoscopic images are to be assessed due to the presence of two different images. This is because the HVS fusion makes the final stereo content perceivable as described in chapter 2, the complete HVS fusion process is not fully comprehended. Hence, other than the quality of the left and the right views and also the introduced disparity between them, the structure of the HVS becomes essential in evaluating the perceived quality of stereoscopic content. Driven both by the entertainment industry and scientific applications in the last decade, an important research topic in IQA, hereafter called *3D QA*, is the quality evaluation of stereoscopic videos. Although 3D QA has been studied



abundantly recently [10,12,17,22,54,60,111,126,130,132,133,168,187], yet it remains relatively unexplored and there is no widely accepted and used objective metric in the research community. However, it is mandatory to evaluate the subjective quality of stereoscopic videos in several test cases and experiments especially when aiming to standardize a new codec targeting 3D content compression [3].

In a special case where asymmetric quality between the views is introduced, it has been shown that the available objective metrics face some ambiguity on how to approximate the perceived quality of asymmetric stereoscopic video [53]. As a result, while in this thesis the asymmetric concept has been exploited frequently in different experiments and studies, subjective evaluation of stereoscopic content becomes an important issue and, hence it will be further explored in section 4.2.

## 4.1 Objective metrics

Objective IQA is accomplished through a mathematical model which is used to evaluate the image or video quality so that it reflects the HVS perception. The goal of such a measure is to estimate the subjective evaluation of the same content as accurately as possible. However, this is quite challenging due to the relatively limited understanding of the HVS and its complex structure as explained in sections 2.2 and 2.3. Yet, considering that the objective metric is a fast and cheap approximation for the visual quality of the content and can be repeated for different processed content easily, it has become a fair substitute of subjective quality assessment in many applications. Therefore, researchers who do not have the resources to conduct systematic subjective tests suffice to report only the objective evaluation of their processing algorithm. However, in several cases e.g. stereoscopic content and especially asymmetric stereoscopic content, subjective tests remain the only trustable option.

The objective quality assessment metrics are traditionally categorized to three classes of full-reference (FRef), reduced-reference (RRef), and no-reference (NRef) [31, 160, 180]. This depends on whether a reference, partial information about a reference, or no reference is available and used in evaluating the quality, respectively.

**FRef metrics** In these metrics, the level of degradation in a test video is measured with respect to the reference which has not been compressed or processed in general. Moreover, it imposes precise temporal and spatial alignment as well as calibration of color and luminance components with the distorted stream. However, in real time video systems, the evaluation with full- and reduced-reference methods are limited since the reference is not available and in most cases no information other than the distorted stream is provided to the metric. Objective quality evaluations reported in this thesis are all using FRef metrics.

**NRef metrics** These metrics mostly make some assumptions about the video content and types of distortion and based on that, try to separate distortions from the content. Since no explicit reference video is needed, this scheme is free from alignment issues and hence, it is not as accurate as FRef metrics.

**RRef metrics** These metrics are a tradeoff between FRef and NRef metrics in terms of availability of the reference information. These metrics extract a number of features from the reference video and perform the comparison only on those features. This approach keeps the amount of reference information manageable in several applications while avoiding some assumptions of NRef metrics.

There exist several different proposals on how to measure the objective quality through automated computational signal processing techniques. In this section several of these metrics are introduced.

The simplest and most popular IQA scheme is the mean squared error (MSE) and Peak-Signal-to-Noise (PSNR) (which is calculated based on MSE). MSE and PSNR are widely used due to the fact that they are simple to calculate, have clear physical meanings, and are mathematically easy to deal with for optimization purposes e.g. MSE is differentiable. However, they have been widely criticized for not correlating well with the perceptual quality of the content, particularly when distortion is not additive in nature [41, 44, 50, 71, 156, 169, 170, 178]. This is expected as MSE is simply the average of the squared pixel differences between the original and distorted images. Hence, targeting automatic evaluation of image quality so that it is HVS-oriented (agrees with the human perceptual judgment), regardless of the distortion type introduced to the content, several other objective measures are proposed [27, 31, 42, 59, 79, 92, 116, 129, 134, 171, 182] and are claimed to correlate more with the HVS perception. Some other well-known objective metrics are briefly explained in the following paragraphs.

**SSIM** Structural Similarity Index [171]. This metric compares local patterns of pixel intensities that have been normalized for the luminance and the contrast. SSIM expression is presented in (4.1) and has been used in [P3].

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4.1)$$

where

$\sigma$  = standard deviation

$\mu$  = mean value of each signal

$C_1$  = constant  $C_1$  is included to avoid instability when  $\mu_x^2 + \mu_y^2$  is very close to 0

$C_2$  = constant  $C_2$  is included to avoid instability when  $\sigma_x^2 + \sigma_y^2$  is very close to 0

**VQM** Video Quality Metric [116]. This metric benefits from several steps. Briefly described, this measure includes 1) sampling of the original and processed video

streams, 2) calibration of both sets of samples, 3) extraction of perception-based features, 4) computation of video quality parameters, and 5) calculation of the general model. The used general model tracks the quality of the perceptual changes presented as distortion in all components of the digital video transmission system (e.g., encoder, digital channel, decoder). There is no simple mathematical way to express the metric, and for more details, the reader is referred to [116].

**PSNR-HVS-M** PSNR Human Visual System Masking [109]. This metric takes into account a model of visual-between contrast masking of the DCT basis functions based on the HVS and the contrast sensitivity function. In this approach first the weighted energy of DCT coefficients for a block with size 8x8 are calculated as shown in (4.2).

$$E_w(X) = \sum_{i=0}^7 \sum_{j=0}^7 X_{ij}^2 C_{ij} \quad (4.2)$$

where

$X_{ij}$  is a DCT coefficient with indices  $i, j$

$C_{ij}$  is a correcting factor determined by the CSF.

However, since the value of masking effect ( $E_w(X)/16$ ) as presented in (4.2) can be too large if an image block belongs to an edge, a new masking effect is proposed in (4.3).

$$E_m(D) = E_w(D)\delta(D)/16 \quad (4.3)$$

where

$$\delta(D) = (V(D1) + V(D2) + V(D3) + V(D4))/4V(D)$$

$V(D)$  = variance of the pixel values in block  $D$

The values of  $C_{ij}$  are calculated as presented in [166].

Now considering the maximal masking effect  $E_{max}$  calculated as  $\max(E_m(X_c), E_m(X_d))$  where  $X_c$  and  $X_d$  are the DCT coefficients of an original and impaired image block, respectively, the visible difference between  $X_c$  and  $X_d$  is determined as (4.4).

$$X_{\Delta ij} = \begin{cases} X_{eij} - X_{dij} & , i = 0, j = 0 \\ 0 & , |X_{eij} - X_{dij}| \leq E_{norm}/C_{ij} \\ X_{eij} - X_{dij} - E_{norm}/C_{ij} & , X_{eij} - X_{dij} > E_{norm}/C_{ij} \\ X_{eij} - X_{dij} + E_{norm}/C_{ij} & , otherwise \end{cases} \quad (4.4)$$

where  $E_{norm}$  is  $\sqrt{E_{max}/64}$

**PSNR-HVS** PSNR Human Visual System [42]. This measure is based on PSNR and universal quality index (UQI) [169] which has been modified to take into account the HVS properties. The modification considers removing the mean shifting and

the contrast stretching using a scanning window according to the method described in [169]. Moreover, MSE is calculated taking into account the HVS according to the approach described in [169]. This is done by first removing the mean shifting and the contrast stretching using a scanning window according to the method described in [169]. Then modified PSNR is defined as in (4.5).

$$PSNR - H = 10 \log \left( \frac{255^2}{MSE_H} \right) \quad (4.5)$$

where  $MSE_H$  is calculated taking into account the HVS according to the approach described in [97] and shown in (4.6).

$$MSE_H = K \sum_{i=1}^{I-7} \sum_{j=1}^{J-7} \sum_{m=1}^8 \sum_{n=1}^8 ((X[m, n]_{ij} - X[m, n]_{ij}^e) T_c[m, n])^2 \quad (4.6)$$

where

I, J denote image size

$K = 1/[(I - 7)(J - 7) \times 64]$

$X_{ij}$  are DCT coefficients of 8x8 blocks

$X_{ij}^e$  are DCT coefficients of the corresponding block in original image

$T_c$  is the matrix of the correcting factors

**VSNR** Visual Signal-to-Noise Ratio [27]. This metric operates via a two-stage approach. In the first stage, contrast thresholds for detecting distortions in the presence of natural images are computed. If the distortions are below this threshold, the image is dimmed to have perfect visual fidelity and no further analysis is required. However, if the distortions are higher than the threshold, a second stage based on the low-level visual property of the perceived contrast and the mid-level visual property of the global precedence is applied. These two properties are modeled as euclidean distances in the distortion-contrast space and VSNR is computed based on a linear sum of these distances. For mathematical expressions the reader is referred to [27].

**WSNR** Weighted Signal-to-Noise Ratio [36]. In this metric, a degraded image is considered as an original image that has been subject to linear frequency distortion and additive noise injection. Then, these distortions are decoupled and the effect of the frequency distortion and the noise quality degradation are calculated via the distortion measure (DM) and the noise quality measure (NQM), respectively. The NQM is based on Peli's contrast pyramid and DM follows three steps of 1) finding the frequency distortion, 2) computing the deviation of that frequency from an all-pass response of unity gains, and 3) weighting the deviation by a model of the frequency response of the HVS. Briefly said, WSNR is defined as the ratio of the average weighted signal power to the average weighted noise power. Since the mathematical way to express the metric is complicated, the reader is referred to [36] for further information.

**VIF** Visual Information Fidelity [134]. This metric quantifies the loss of image information during the degradation process and explores the relationship between the image information and the visual quality. The model calculates the information that is presented in the reference image and based on how much of this reference information can be extracted from the distorted image, the subjective quality of the processed image is estimated. For respective equations reader is referred to [134].

**MS-SSIM** Multi-Scale Structural Similarity Index [172]. This method considers the assumption that the HVS is highly adapted for extracting structural information from the scene. Therefore, the proposed method is a multi-scale structural similarity method (more flexible than single scale methods) exploiting an image synthesis algorithm to calibrate the parameters that define the relative importance of different scales. This is briefly described in (4.7).

$$SSIM(x, y) = [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^M [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \quad (4.7)$$

where

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4.8)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (4.9)$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (4.10)$$

where

$C_1$ ,  $C_2$ , and  $C_3$  are small constants similar to those introduced in (4.1)

$\alpha_M$ ,  $\beta_j$ , and  $\gamma_j$  are used to adjust the relative importance of different components

All metrics listed above, except VQM, are computed on the luma component of the frame and the final index value for the whole sequence is averaged across the results of frames.

The accuracy of PSNR and some other objective quality metrics to measure the subjective quality has been studied recently with stereoscopic viewing [56, 57]. While no perfect correlation between any objective metric and the subjective results were found, PSNR and some other FRef objective metrics were found to provide a reasonable correlation with subjective ratings. Since there were no drastic differences between different objective metrics, other than the conducted subjective experiments in this thesis, PSNR or SSIM has been utilized as the objective quality evaluation experiments in several publications [P1], [P3], [P4], [P5], and [P8].

---

## 4.2 Subjective quality assessment

The subjective video quality assessment methods are based on one or several groups of naïve or trained subjects viewing the video content, and scoring the quality of the shown content [115]. Moreover, these tests should meet the ITU-T recommendations for subjective quality assessment and hence, the tests must follow strict conditions e.g. room illumination, viewing distance, test duration, content presentation, and evaluators' selection [114]. However, as subjective tests involve a separate rating for each and every stimulus by all users, it is in general quite time consuming, especially for a large set of test material. Therefore, it is not always possible to conduct such a test and mostly it is limited to the cases where a reliable and vital decision needs to be made. Considering the duration of the test, it cannot be used in cases where a fast judgment needs to be made, e.g. the decisions which are made at the encoder to optimize the coding parameters. Despite these drawbacks, as the results are quite precise, subjective quality assessment is the most common approach used for formal quality evaluation [3].

As discussed previously in 3D QA, since the HVS fusion is involved, the best method is subjective assessment. Moreover, if asymmetry is used in the left and right view, the objective evaluation schemes perform even worse in estimating the subjective quality. Hence, in this thesis for several research topics [P2], [P3], [P6], [P7], [P8], [P9], and [P10], subjective test evaluation were carried out to obtain reliable quality assessment of stereoscopic content.

### 4.2.1 Test procedure

There are several methodologies to conduct a subjective test [106] and the international recommendations for subjective video quality assessment e.g. ITU-R BT.500-11 [115] specify how to perform different types of subjective quality assessments. In general, these tests can be divided into two types, namely Single Stimulus Impairment Scale (SSIS) and Double Stimulus Impairment Scale (DSIS) [115]. In SSIS, viewers only evaluate the quality of impaired stimulus i.e. observing only the processed content and rating them. On the other hand, in DSIS there is a reference where subjects rate the quality or change in the quality while switching from the reference to the video content being evaluated. In this approach, each impaired video is accompanied with its reference video so the subjects can always evaluate the respective quality of the processed content compared to the anchor. Each approach has its own advantages e.g. DSIS method is claimed to be less influenced by content. In other words, ratings in DSIS, evaluation is less sensitive to the quality drop level and also presentation order of impaired content as they are constantly compared to the reference and hence, are evaluated more robustly. On the other hand, it has been claimed that a more representative quality estimate for quality monitoring applications results by considering SSIS [106]. DSIS [115] has been used in [P2], [P3], [P6], [P7], [P8], [P9], and [P10].

Subjects attending each subjective test can be selected from naïve or expert/experienced users, depending on the target of the quality investigation. In most cases, un-experienced users in the video coding field are selected, as mostly the output of the encoding is to be observed by all types of subjects and is not limited to a specific group of users. However, in some cases the target of an experiment is to assess the presence and severity of some specific coding artifact, e.g. blurring or blocking artifacts, and hence, experts are selected as subjects to evaluate the subjective quality.

Prior to performing each subjective test, several visual tests should be conducted for all participants to confirm that they are eligible to attend the test. Specifically, subjects are first tested for far and near visual acuity, stereoscopic acuity (Randot test), contrast sensitivity (Functional Acuity Contrast Test), horizontal and vertical phoria (Maddox wing test [120]), near point of accommodation and convergence RAF gauge test [95], and the interpupillary distance. For stereoscopic quality evaluation, stereo vision is also evaluated to confirm that the subjects are capable of evaluating 3D content. In all experiments, viewers who are not found to have normal visual acuity and stereopsis are rejected. Different prior visual evaluations are used and presented in different experiments presented in this thesis. Moreover, at the beginning of each test we started the evaluation with a combination of anchoring and training. Participants were shown both extremes of the quality range of the stimuli to familiarize them with the test task, the contents, and the variation in quality they could expect in the actual tests that followed. We presented each video sequence twice and in a random order for all test experiments to achieve more accurate scores from each subject.

## 4.2.2 Analyzing subjective scores

In subjective evaluations, conventionally the results are reported based on the average subjective scores and the 95% Confidence Interval (CI). The 95% CI reports the interval in which 95% of the scores are located and hence, is a good representative on how close the subjective scores were. So, if the relative length of CI is small compared to the used scale, it shows accuracy of quality estimation from subjects and increases the reliability of the results. The amount of overlap between confidence intervals of subjective scores for two test cases is an indicator whether they are significantly different or not. However, it is not always possible to evaluate the significance difference between the quality of two test cases based on subjective score figures and further mathematical analysis is required. This concept is handled typically by considering the raw scores from different subjects and different test cases and making the comparison over those values rather than only considering the mean score and 95% CI for each test case. Wilcoxon's test [175] is one method to measure differences between two related and ordinal data sets as used in [34]. Conventionally a significance difference level of  $p \leq 0.05$  is used to separate the subjective scores for different test cases, i.e. when  $p > 0.05$  then the test cases are supposed not to

have subjectively significant difference; otherwise, their subjective quality is considered distinguishable. In publications [P2], [P6], and [P8] included in this thesis such analysis was used to make statistical conclusions about the preference of subjects.

## 4.3 Subjective quality of 3D video

In the recent years, an increase in the number of 3D movies and applications, e.g. 3D gaming and hand-held devices featuring a 3D display, is observed. Moreover, few television channels are commercially broadcasting stereoscopic video content while several user devices are already capable of processing stereoscopic content. One of the principal methods to extract 3D content is to watch the content with shutter glasses or polarized glasses (as explained in section 3.3). However, there are several cases where 3D content is playing on a display but viewers are not necessarily using viewing glasses and cannot be considered as active 3D viewers. The recently opened Sky 3D pubs in Ireland and UK [4] are examples of such situation where some costumers are active 3D viewers while others, e.g. the staff or other costumers are just momentarily peeking at the display showing 3D content. To evaluate the subjective quality of a stereoscopic video, we consider two cases where in the first case, users wear viewing glasses and in the second case, they watch the same content without viewing glasses. These are discussed in the following sub-sections.

### 4.3.1 Viewing 3D content with glasses

Subjects observing the stereoscopic content with glasses expect good general quality of the content (i.e. no encoding artifacts) as well as an acceptable depth sensation. High quality of the content can be achieved with an optimized encoding of the views considering the available budget for the bitstream while depth perception can be handled based on the disparity between the views. In other words, increasing the disparity between the left and right views increases the depth perception while decreasing the disparity causes less depth feeling and hence, lower 3D sensation.

An annoying artifact while watching 3D content with glasses is cross-talk [70]. This artifact is perceived as shadow or double contours due to imperfect optical separation between the left and the right views by filters of passive glasses or small lack of synchronization between shutters on active glasses and the displayed left and right views on the display [144]. Under this condition, the opposite view, which should have been blocked by the viewing glasses, is observed by the non-respective eye causing a cross-talk. This has been reported as one of the main disturbing perceptual effects while watching stereoscopic content with viewing glasses [65]. The extreme case of such an artifact happens in the case when viewing glasses are not used to observe the 3D content. In this condition, both the left and right views are equally visible to both eyes and the subjective quality of the content becomes unpleasant and not acceptable. This topic is further covered in the next sub-section.



### 4.3.2 Viewing 3D content without glasses

3D video observation on stereoscopic display without glasses sounds quite annoying since each eye sees both views simultaneously. This effect is called *ghosting artifact* in this thesis where one view appears as a ghost besides the other view and since both eyes observe both views simultaneously and with the same quality and intensity, a double edge effect due to ghosting artifact is observed. In this case, the 3D perception can be improved by decreasing the visibility of one view and increasing the similarity of its presentation to the other view. This can be done through some modifications and renderings applied to one of the views taking into account the characteristics of both views and the stereoscopic video in general. Hence, the ghosting artifact is reduced and subjects can view a more pleasant content which is modified to present a more aligned content with one of the two views as in 2D presentation. However, it should be noted that this rendering should not sacrifice the depth sensation of the same content when viewed with glasses.

A new algorithm has been introduced in [P2] to render two views in such a way that one view is selected as the dominant view while the other is marked as the non-dominant view. The non-dominant view was modified to become more similar to the dominant view and a threshold between the 2D and 3D presentation was found through a series of subjective tests. This threshold defines how and to what

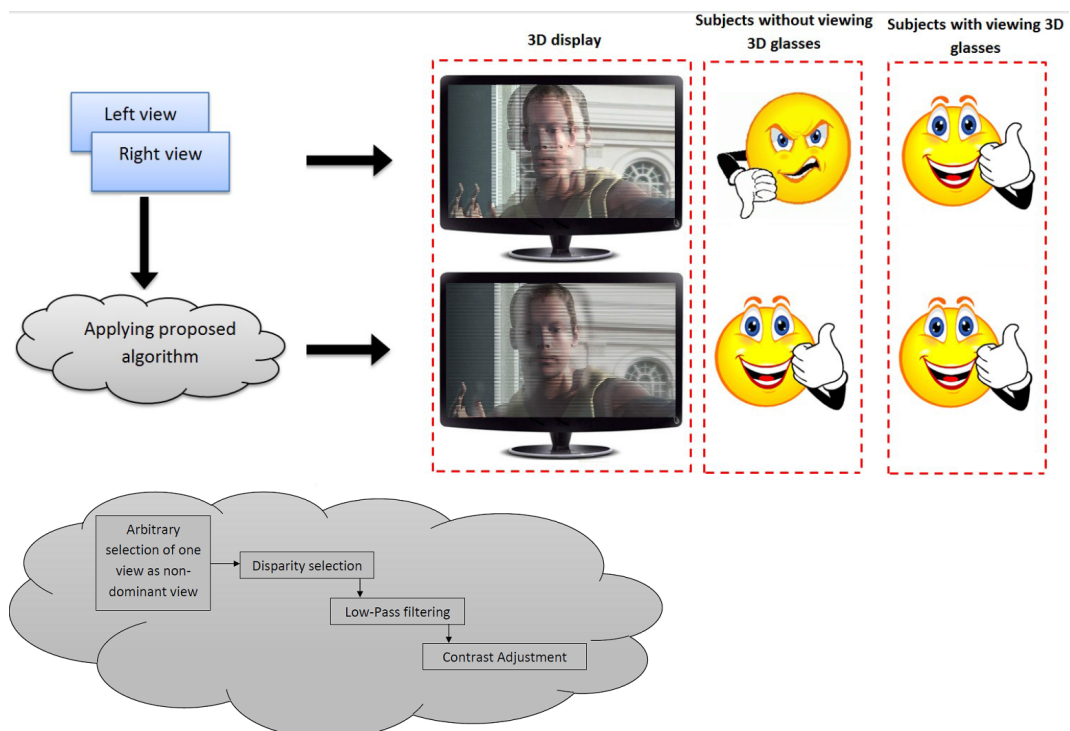


Figure 4.1: Simultaneous 2D and 3D presentation of 3D content as introduced in [P2]

---

extent the non-dominant view can be modified so that both 2D and 3D perceived qualities remain satisfactory. The novel technique is represented in Figure 4.1, while Figure 4.2 presents few sample images comparing the 2D presentation of the original and modified stereoscopic content with the algorithm proposed in [P2]. In these images the manipulated stereopairs tend to be more similar to 2D views while the 3D perception and depth feeling is preserved to an acceptable extent. However, in the 2D presentation, the relative increase in the subjective quality of stereoscopic content rendered with the proposed algorithm compared to the quality of the original stereoscopic content is considerably higher when shown on a 3D display compared to what is presented in Figure 4.2.



Figure 4.2: 2D presentation of stereoscopic video combinations from (a) original stereopair and (b) proposed rendered stereopair

# Chapter 5

## Asymmetric Stereoscopic Video

### 5.1 Introduction

In stereoscopic videos two synchronized, monoscopic video streams are included and normally the left and the right views have similar quality, i.e., both views have the same spatial resolution and have been identically encoded. In some cases, the quality of one view is intentionally degraded compared to the other one. This is attributed to the widely believed assumption of the binocular suppression theory [15] that the HVS fuses the two images in such a way that the perceived quality is closer to that of the higher quality view. The general presentation of an asymmetric stereoscopic video is depicted in Figure 5.1 where the encoding has decreased the quality of the left view compared to the right view.

Considering that asymmetric stereoscopic video deals with quality reduction in one view, one might ask what are the criteria to decide which view should have the lower quality and on which factors does this decision depend. One important related



(a) Lower quality view

(b) Higher quality view

Figure 5.1: Asymmetric stereoscopic video

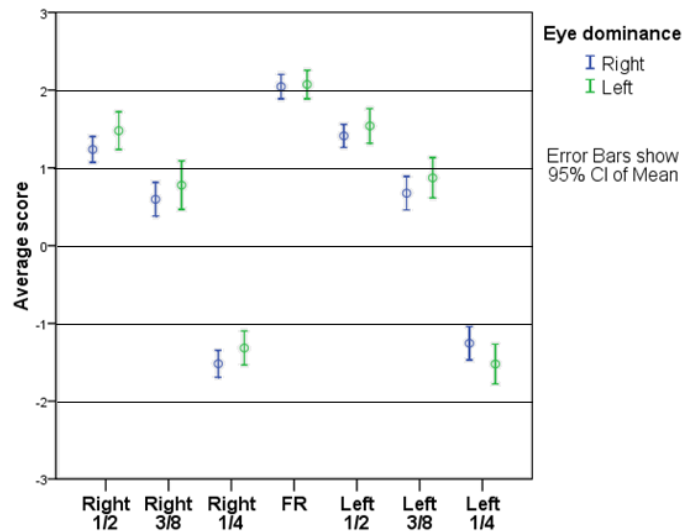


Figure 5.2: Average subjective ratings and 95% confidence intervals for different eye dominant subjects

topic is the eye dominance of viewers. Approximately 70% of the world population is right eye dominant [110] while the rest are left eye dominant. This assures that if eye dominance has an effect on the perceived quality of asymmetric video, it is better to provide the left view with a lower quality, since the majority of viewers are right eye dominant. One approach to test this is to conduct an experiment with asymmetric stereoscopic sequences where in half of the sequences, the left view has a lower quality, while in the other half, the right view has a lower quality. A group of left and right eye dominant subjects are then asked to view these sequences and rank them subjectively. The relevance of eye dominance can then be evaluated for asymmetric stereoscopic video. In [8] such an experiment was performed and it was discovered that subjective ratings of asymmetric quality sequences are not statistically impacted by eye dominance. This conclusion was achieved for different video sequences while the same outcome has also been confirmed in the literatures, e.g., [88,130] for still images. The experiments included using MR stereoscopic video where in half of the test material the left view was downsampled with ratios  $\frac{1}{2}$ ,  $\frac{3}{8}$ , and  $\frac{1}{4}$  in the horizontal and vertical directions and in the other half, the right view was downsampled with the same ratios. Moreover, half of the subjects were right eye dominant while the other half was left eye dominant. All subjects assessed the quality of both asymmetric video sequences. The results of the mean subjective scores and the associated 95% CI are depicted in Figure 5.2. Considering very close mean subjective scores and largely overlapping 95% CI, it was concluded that eye dominance of the viewers and the perceived subjective quality based on the direction of asymmetry (left view or right view having a lower quality) are not statistically related.

---

## 5.2 Types of asymmetry

There are several approaches to achieve asymmetry between the two views of a stereoscopic video. Different techniques are depicted in Figure 5.3, while two or more methods can be combined and used to obtain asymmetric videos.

**Mixed-resolution (MR) stereoscopic video.** MR was first introduced in [105] and is also referred to as resolution-asymmetric stereoscopic video. In this scheme, one of the views is low-pass filtered and hence has a smaller amount of spatial details. Furthermore, the low-pass filtered view is usually sampled with a coarser sampling grid, i.e., represented by fewer pixels. This view should be upsampled to the FR i.e. the same resolution as the other view before being displayed. Therefore, the combination of LPF and subsampling causes blurring effect on the manipulated view.

**Mixed-resolution chroma sampling.** The chroma pictures of one view are represented by fewer samples than the respective chroma pictures of the other view [11]. It has been shown [11] that downsampling chroma components does not affect the subjective quality of asymmetric video compared to symmetric video provided that the downsampling ratio is properly selected, as it depends on the image size and the image content.

**Asymmetric sample-domain quantization.** The sample values of the two views are quantized with a different step size [P9]. For example, the luma samples of one view may be represented with the range of 0 to 255 (i.e., 8 bits per sample) while the range may be scaled to 0 to 159 for the second view. Thanks to fewer quantization steps, the second view can be compressed with a higher ratio compared to the first view. Different quantization step sizes may be used for luma and chroma samples. As a special case of asymmetric sample-domain quantization, one can refer to bit-depth-asymmetric stereoscopic video when the number of quantization steps in each view matches a power of two.

**Asymmetric transform-domain quantization.** The transform coefficients of the two views are quantized with a different step size. As a result, one of the views has a lower fidelity and may be subject to a greater amount of visible coding artifacts, such as blocking and ringing. This type of asymmetry has been studied in [P3] and [130].

**A combination of different encoding techniques.** An example of such a combination is illustrated in 5.3.e. This technique combines MR and asymmetric transform domain quantization was explored further in [P3], [P9], and [20, 152].

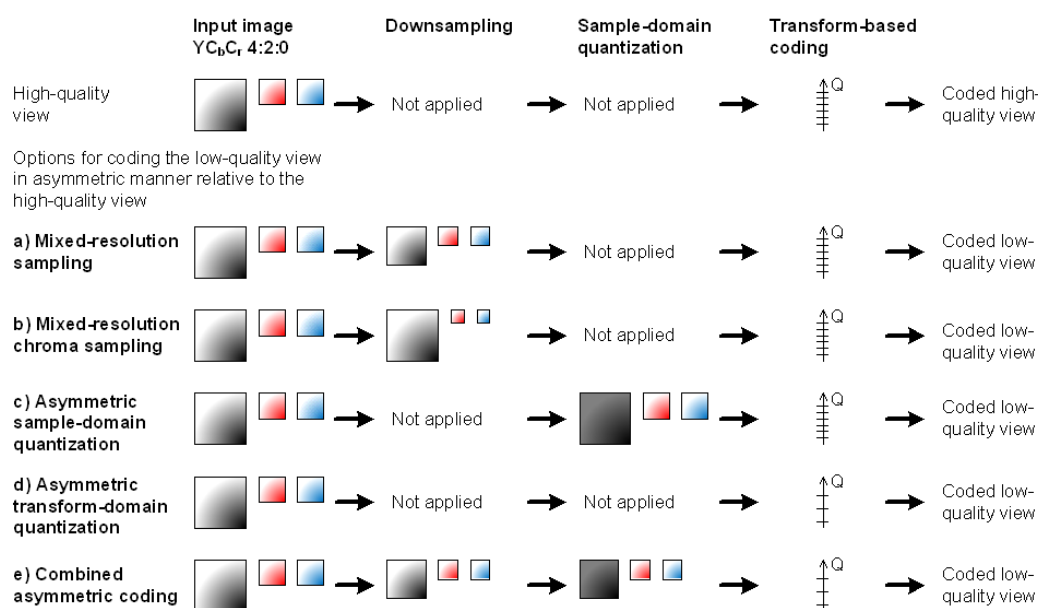


Figure 5.3: Examples of different types of asymmetric stereoscopic video coding

Each method brings some efficiency and higher performance while potential drawbacks are introduced to the codec too. In the next section the performance of each asymmetric approach is analyzed in more detail by presenting the comparison results with the anchor symmetric schemes.

## 5.3 Motivation for using asymmetric stereoscopic video

In this section, we go through different types of asymmetric schemes introduced in the previous section and justify their utilization. First a general introduction to low-pass filtering and down/upsampling processes is provided, as they are commonly used in some of the proposed methods.

### 5.3.1 Low-pass filtering

Low-pass filtering the texture views targets removing the high frequency components while keeping the spatial resolution and general structure of the image untouched. This enables the compression of the same content with reduced number of bits since less detail (high frequency components) need to be encoded. In the case where videos are presented in polarized displays, a downsampling with ratio  $\frac{1}{2}$  along the vertical direction is applied to the content. This is because the vertical spatial resolution of the display is divided between the left and right view and hence, each one has half the vertical resolution (as described in sub-section 3.3.1). In such cases, depending on the display and content, a huge aliasing artifact [33] might be introduced while

perceiving the stereoscopic content. However, applying LPF reduces such artifact considerably since the high frequency components responsible for the creation of aliasing are removed in a pre-processing stage [52]. Therefore, in several cases such as [1], low-pass filtering is applied before spatial downsampling the image to reduce aliasing. The down/up sampling algorithms are further discussed in the next subsection.

### 5.3.2 Down/up sampling

Downsampling or subsampling in signal processing reduces the sampling rate of a signal. This is usually done to reduce the data rate or the size of the data. Image downsampling is performed by selecting a specific number of pixels, based on the downsampling ratio, out of the total number of pixels in the original image. This will result in presenting the original image with a lower respective spatial resolution.

Downsampling in image/video coding is basically applied in one of the first steps to reduce the complexity of the next steps introduced in the coding scheme. The complexity reduction refers to a decrease in number of operations per pixel required to encode the original image. Hence, if an encoding algorithm is used to compress an image with size  $W$  (width) and  $H$  (height), after downsampling the image to resolution  $W'$  and  $H'$  (where  $W' \leq W$  and  $H' \leq H$ ), the required number of operation per pixel are kept constant while the total number of operations required to encode the downsampled image are reduced by factor  $\frac{W'}{W} \times \frac{H'}{H}$ . This is an important issue in several cases where a limited power is available for the codec, e.g. in mobile or other handheld devices which work with battery. Figure 5.4 illustrates the execution time required to encode one view with full and quarter resolution for several sequences with 3DV-ATM software [5]. The simulations were performed on Windows OS with a quad-core CPU with a clock rate of 2.8GHz. These results are in agreement with those reported in our experiments in [P3] and show a substantial execution time reduction while encoding the lower resolution content. Furthermore, by downsampling it is possible to reduce the needed storage memory and bandwidth required for transmission and/or broadcasting. After storage, manipulation, or transmission, the downsampled content will be upsampled to the original resolution for displaying.

Other than downsampling ratio, the format of the video should be considered while downsampling too. Depending on whether it has e.g. YUV or RGB format, downsampling should be applied to different components in such a way that the outcome is well presented by the same format. Simple subsampling is the easiest method to downsample a video by selecting one pixel value to represent a group of pixels. There are numerous filters proposed and analyzed by researchers [55,93,128,183], to perform the required down/up sampling but the filters presented in [1] are the standardized approach utilized widely by researchers. In this approach the anti-aliasing filter will be applied before downsampling and upsampling is based on interpolating the missing pixel values by a 6-tap filter. This state-of-the-art resampling technique is designed specifically for best perceived quality of video after down/up sampling



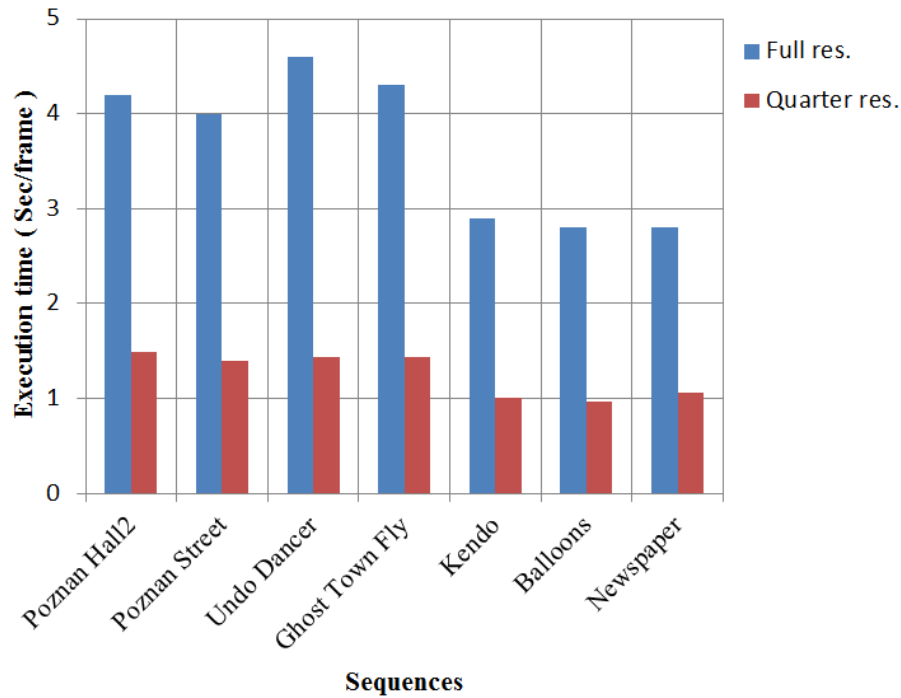


Figure 5.4: Encoding times for full and quarter resolution views

and provides a higher PSNR for subsampled texture views compared to other similar proposed algorithms.

### 5.3.3 Performance analysis of different asymmetric types

In this sub-section, we review the efficiency of different asymmetric stereoscopic types and report the results and previous art concerning each scheme.

**Mixed-resolution (MR) stereoscopic video coding** This is one of the commonly used and well-studied types of asymmetry between the views. A major force behind many research activities in video coding is to reduce the complexity of the straight-forward encoder and decoder implementation as decreasing the spatial resolution of one view results in reducing the number of pixels involved in the encoding and decoding compared to the case where FR content is used. In return, two steps are added to the whole process from encoding the original input video until displaying the content for the end user. These steps are downsampling the original FR frame using the associated downsampling ratio prior to encoding and upsampling the decoded frame to have FR frames in both views of the final stereoscopic video. This is depicted in Figure 5.5. However, the downsampling and upsampling perform considerably smaller number of operations per pixel compared to encoding and decoding and hence, is expected to yield substantial reduction in complexity [P3]

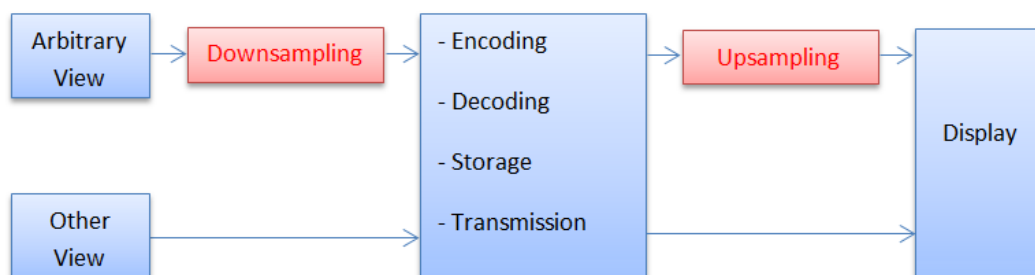


Figure 5.5: Block diagram illustrating the placement of down and upsampling blocks for different applications

and [20, 43].

Another benefit of MR stereoscopic video coding is the bitrate reduction due to the smaller number of pixels to be encoded compared to the FR case. If the left and right views are encoded in simulcast mode (no inter-view prediction) the bitrate needed to encode MR with the same quantization parameter (QP) as FR stereoscopic video is reduced as a smaller number of pixels is encoded. The amount of bitrate reduction depends on the downsampling ratio and video content. However, this comes at the price of degrading the subjective quality of the view with the lower spatial resolution. The subjective quality of MR scheme has been extensively studied in the literature [P5], [P6], [P10], and [20, 142, 152]. The results confirm that the perceived quality of the MR videos is closer to that of the higher resolution view.

The subjective impact of uncompressed MR sequences at downsampling ratios of  $\frac{1}{2}$  and  $\frac{1}{4}$  applied both horizontally and vertically was studied in [142]. A combination of a data projector and shutter glasses were used as the viewing equipment with a viewing distance equal to  $4H$  (where  $H$  is the height of the frame). It was found that the perceived sharpness and the subjective image quality of the MR image sequences were nearly transparent at the downsampling ratio of  $\frac{1}{2}$  in both directions but dropped slightly at the ratio of  $\frac{1}{4}$ .

In [152], it was confirmed that the perceived quality of MR video was closer to the subjective quality of the view with the higher resolution. In this thesis also a series of subjective tests was performed to evaluate the perceived quality of compressed MR stereo video compared to FR stereo video. Results in [P3] showed that in most cases, if one view is downsampled with a ratio of  $\frac{1}{2}$  along both coordinate axes, the subjective quality will not degrade considerably compared to FR scheme, under the same bitrate constraint. In addition to confirming the results of [152], conclusions in [P3] reveal that most compressed MR video sequences where one view is downsampled with a ratio  $\frac{1}{2}$  provide a similar subjective quality to FR scheme, while decreasing this ratio to  $\frac{3}{8}$  introduces severe quality degradation that rejects the idea of exploiting such downsampling ratio in MR format.

To increase the coding performance, and as introduced in H.264/MVC [29], inter-view prediction can be enabled. An implementation of MR scheme includ-

ing inter-view prediction enabled is presented in [20]. However, in this case, since the spatial resolution of the left and right views is not the same, the performance of inter-view prediction is lower compared to FR scheme where both views have the same resolution. Authors in [20] performed two sets of subjective studies for full- and mixed-resolution stereo video on a 32-inch polarization stereo display and on a 3.5-inch mobile device. In MR scheme, the spatial resolution of one view was downsampled to half in both directions. The results revealed that the higher the resolution, the smaller the subjective difference is between FR and MR stereoscopic video. An equivalent result was also discovered as a function of the viewing distance by changing the distance from 1 to 3 meters. The conclusion was that the greater the viewing distance, the smaller the subjective difference becomes between FR and MR. Moreover, the study showed that the performance of the encoding process differed based on the direction of the inter-view prediction. It was shown that the prediction from the high resolution to the low resolution view, outperforms the prediction from the low to the high resolution view.

Asymmetry achieved with MR scheme does not always have to include downsampling one view while keeping the other view with FR. In this thesis, it has been shown that downsampling different views along different directions may result in a better subjective quality compared to the conventional MR schemes [P6]. This scheme, called *cross-asymmetric MR*, considers the SI and characteristics of each view and chooses the direction in which each view should be downsampled and hence, one view is downsampled in vertical direction while the other view is downsampled in horizontal direction. The subjective results [P6] show that this scheme outperforms conventional MR scheme and this is because of performing automatic downsampling based on the content of each view and preserving the spatial resolution of views in the directions where they have the higher SI. Moreover, the number of pixels involved in the encoding and decoding process decreases in the proposed scheme.

Another research conducted in this thesis based on the principle of MR asymmetric texture is introduced in [P5] wherein a depth-enhanced multiview scenario including 3 views, the spatial resolution of the side views is reduced to quarter the resolution of the central view and hence, an average 4% and 14.5% delta bitrate reduction (using Bjontegaard delta bitrate and delta luma PSNR metrics [14]) for coded and synthesized views is achieved, respectively. This topic is further discussed in sub-section 6.6.2.

In general, it can be concluded that MR stereoscopic video is a promising approach to decrease the bitrate and complexity and yet achieve comparable quality compared to FR scheme. However, the downsampling ratio and the type of MR scheme should be selected based on the targeted application and the video content to provide the highest efficiency.

**Mixed-resolution chroma sampling.** Changing the spatial resolution of the chroma component was already discussed with the MR stereoscopic video. However, [11]

perform analysis on stereo images and reports that if downsampling is only applied to chroma components, the subjective quality of the decoded data is not degraded much on a stereoscopic display. This approach also benefits from the lower bitrate consumption for the encoding and also the complexity decrease both at the encoder and decoder.

**Asymmetric sample-domain quantization.** In this approach, the pixel values of the left and right views are quantized utilizing a different quantization step size [P9]. This is done by changing the scaling range e.g. following the same algorithm used for the weighted prediction mode of the H.264/AVC standard [117]. This is reported in (5.1):

$$q = \text{round}\left(\frac{i \times w}{2^d}\right) = (i \times w + 2^{d-1}) \gg d \quad (5.1)$$

where:

$q$  is the quantized sample value

$\text{round}$  is a function returning the closest integer

$i$  is the input value of the luma sample

$w$  is the explicit integer weight ranging from 1 to 127

$d$  is the base 2 logarithm of the denominator for weighting (fixed to 8 in our experiments)

This equation is the same formula used in H.264/AVC weighted prediction and  $\frac{w}{2^d}$  is referred to as the luma value quantization ratio.

Inverse quantization of sample values to their original value range is achieved by (5.2):

$$r = \text{round}\left(q' \times \frac{2^d}{w}\right) \quad (5.2)$$

where:

$r$  is the inverse-quantized output value

$q'$  is the scaled value of the luma sample as output by the transform-based decoder

Other parameters are the same values as used in the sample value quantization (5.1).

Applying such quantization prior to encoding guarantees a relatively lower bitrate compared to the case where quantization is not applied. However, a tradeoff between the subjective quality degradation and the bitrate reduction should be considered when exploiting this type of asymmetry.

Considering our scheme presented in [P9], we studied in which conditions MR stereoscopic video coding outperforms symmetric stereoscopic video coding. The results were presented over both MR coding and MR coding applied together with asymmetric sample-domain quantization. These results were reported in [P9]; however, here the conclusions of those results are further analyzed statistically.

Table 5.1: Spatial resolution of the sequences for different downsampling rates

	Full	$\frac{5}{6}$	$\frac{3}{4}$	$\frac{1}{2}$
Undo Dancer	960x576	800x480	720x432	480x288
Others	768x576	640x480	576x432	384x288

The simulations were performed with four sequences: Undo Dancer, Kendo, Newspaper, and Pantomime. A display capable of standard definition (SD) television or wide SD was the target display in these experiments. Hence, the sequences were downsampled from their original resolutions to the lower resolution (Full) as mentioned in Table 5.1.

For each sequence, the left view was coded using H.264/AVC [117] while three pre/post processing methods i.e. downsampling, sample value quantization, and transform coefficient quantization, were applied to the right view, which was also coded with H.264/AVC. The comparison was made so that the bitrate of the left and the right views for different combinations was always kept the same. The coding methods included in the subjective comparison were the following:

1. Symmetric stereoscopic video coding. No downsampling or quantization of luma sample values.
2. MR stereoscopic video coding.
3. Combined MR and asymmetric sample-domain quantization.

In order to have a representative set of options for MR coding, three bitstreams per sequence and bitrate were generated, each having a different downsampling ratio for the lower-resolution view. The subjective results achieved for stereoscopic video in [8] motivated us to use downsampling ratios equal to or greater than  $\frac{1}{2}$ . Hence, downsampling was applied to obtain a spatial resolution of  $\frac{1}{2}$ ,  $\frac{3}{4}$ , and  $\frac{5}{6}$  relative to the FR along both coordinate axes. Table 5.1 presents the spatial resolution used for different sequences.

As the number of potentially useful combinations for the downsampling ratio and the luma value quantization ratio is large, their joint impact on the subjective quality was studied first through expert viewing to select particular values for the downsampling ratio and the luma value quantization ratio for the subsequent formal subjective quality evaluation. The following subset of asymmetric parameter combinations was found to be performing well and hence selected to be tested systematically:

1. MR stereoscopic video coding, downsampling ratio  $\frac{1}{2}$
2. MR stereoscopic video coding, downsampling ratio  $\frac{3}{4}$
3. MR stereoscopic video coding, downsampling ratio  $\frac{5}{6}$
4. MR stereoscopic video coding, downsampling ratio  $\frac{3}{4}$ , combined with asymmetric sample-domain quantization with ratio  $\frac{5}{8}$  i.e.  $d = 3$  and  $w = 5$ .

In order to compare the selected coding schemes, bitstreams with an equal bitrate were generated. In order to keep the duration of the subjective viewing session

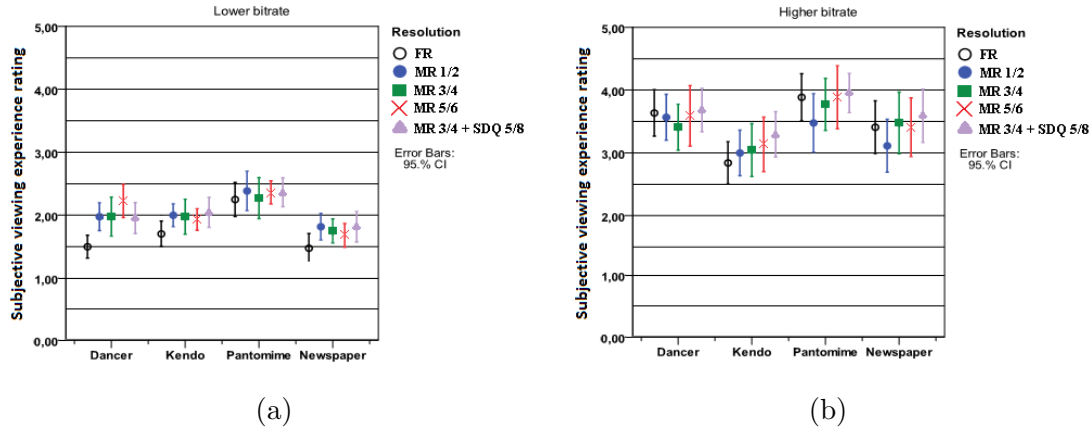


Figure 5.6: Subjective test results for (a) low bitrate and (b) high bitrate sequences

reasonable, only two bitrates were selected and used for the formal subjective test experiment. The QP selection of different methods is reported in Table 5.2 while the tested bitrates are presented in Table 5.3. Moreover, Table 5.3 includes the PSNR values that were achieved with symmetric coding in order to provide a rough quality characterization of the tested sequences.

12 subjects attended this experiment. Their age varied from 19 to 32 years with an average age of 23.6 years. Figure 5.6 shows the average subjective viewing

Table 5.2: QP selection of different methods for the left view (right views are identical for different coding methods of each sequence)

Resolution	QP of Lower - <b>Higher</b> bitrate				
	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{5}{6}$	$\frac{3}{4}$
Sample-domain quantization	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{5}{8}$
Pantomime	44 - <b>35</b>	35 - <b>28</b>	40 - <b>32</b>	41 - <b>33</b>	36 - <b>30</b>
Undo dancer	45 - <b>32</b>	35 - <b>24</b>	40 - <b>28</b>	42 - <b>30</b>	36 - <b>26</b>
Kendo	45 - <b>38</b>	34 - <b>29</b>	40 - <b>34</b>	42 - <b>35</b>	36 - <b>30</b>
Newspaper	45 - <b>33</b>	35 - <b>26</b>	40 - <b>30</b>	42 - <b>31</b>	36 - <b>26</b>

Table 5.3: Tested bitrate values per view and the respective PSNR values achieved by symmetric stereoscopic video coding with H.264/AVC

Sequence	Bitrate (Kbps) - PSNR (dB)	
Pantomime	445.8 - <b>31.93</b>	343.9 - <b>30.0</b>
Undo dancer	301.5 - <b>29.2</b>	224.6 - <b>27.73</b>
Kendo	280.3 - <b>33.25</b>	238.5 - <b>32.0</b>
Newspaper	148.0 - <b>30.0</b>	115.4 - <b>28.3</b>

Table 5.4: Statistical significance differences (SSD) of asymmetric methods against FR symmetric(1 = there is SSD, 0 = No SSD)

Quality	Asymmetric coding	Sequence			
		Undo Dancer	Kendo	Pantomime	Newspaper
Lower bitrate	MR $\frac{1}{2}$	1	1	0	1
	MR $\frac{3}{4}$	1	1	0	1
	MR $\frac{5}{6}$	1	1	0	0
	MR $\frac{3}{4}$ + SDQ $\frac{5}{8}$	1	1	0	1
Higher bitrate	MR $\frac{1}{2}$	0	0	0	0
	MR $\frac{3}{4}$	0	0	0	0
	MR $\frac{5}{6}$	0	0	0	0
	MR $\frac{3}{4}$ + SDQ $\frac{5}{8}$	0	1	0	0

experience ratings for all bitstreams. It can be concluded from Figure 5.6a that the asymmetric subjective results outperformed the FR symmetric approach in 3 out of 4 cases in the lower bitrate. On the other hand, Figure 5.6b suggests that at a higher bitrate, no asymmetric coding method significantly outperformed the FR symmetric case. These observations were confirmed with statistical significance comparison results achieved by the Wilcoxon signed-rank test [175] as presented in Table 5.4. In this flag table, 1 presents statistical significant differences (SSD) between subjective scores while 0 shows no SSD between the ratings. In Table 5.4 all subjective scores of MR schemes is compared against FR scheme while no SSD among the different MR methods was observed. Considering that the quality difference of the lower and higher bitrates was only 1.58 dB in average luma PSNR (see Table 5.3), we believe that there exists a threshold which governs whether the subjective quality dominance switches between symmetric and asymmetric compression methods. This threshold appeared to be sequence dependent as seen from the PSNR values reported in Table 5.3 and hence, should be further studied. Yet, it is an informative indicator on the existence of such fine threshold separating the dominance of symmetric and asymmetric content under tested conditions.

**Asymmetric transform-domain quantization** This is mostly done by applying different quantization steps to transform coefficients of the left and right views. This approach has been extensively studied in the literature [19, 125, 145, 152] and the general conclusion is that the perceived quality of the quality-asymmetric videos is approximately equal to the average of the perceived qualities of the two views. This conclusion has also been confirmed in one experiment presented in [P3] where the subjective scoring of symmetric and quality-asymmetric stereoscopic videos were found to be similar.

**Combining different asymmetric schemes** A presentation of different combinations of asymmetry is illustrated in Figure 5.3e. In this scheme, one view is not manipulated; however, the other view is impaired using more than one processing phase e.g. a combination of spatial resolution reduction, chroma downsampling, increasing the QP applied to the transform coefficients, and/or sample value quantization. The performance of such a combination should be verified on different content as it has been shown that there might exist a threshold which by crossing it, the preference between symmetric and asymmetric stereoscopic content switches. Research results on MR and asymmetric transform domain quantization is presented in [P3], [P9] and [20, 152].

## 5.4 Limits of asymmetry

While confirmed in the literature that different types of asymmetric stereoscopic video in several cases visually outperform symmetric stereoscopic videos [P3], [P6] and [8, 20, 124], the amount of this asymmetry remains ambiguous. Many researchers have done excessive experiments to determine the limits for different types of asymmetry between the left and right view of a stereopair so that the quality decrease is not visible to viewers, and yet a higher encoding performance is achieved compared to the symmetric case.

One general conclusion for the asymmetric subjective quality of stereoscopic video as a function of the viewing distance is that by increasing the viewing distance, the perceived difference between MR and FR stereopair decreases [P10] and [20]. This is due to the fact that the high frequency components removed from one view, due to LPF applied before downsampling the spatial resolution, are not visible from a further distance while they might be more noticeable when the viewer is closer to the display.

In MR asymmetric stereoscopic videos, a downsampling ratio applied to one view has a critical rule in the final subjective quality of the content. In this thesis, the impact of downsampling ratio in MR stereoscopic video was studied [P3]. Downsampling ratios  $\frac{1}{2}$ ,  $\frac{3}{8}$ , and  $\frac{1}{4}$  were applied vertically and horizontally and stereo video sequences were played on a 24-inch polarized display. A correlation comparison between the subjective results and the average luma PSNR showed that under our test condition, there exists a breakdown point between downsampling with ratios  $\frac{1}{2}$  and  $\frac{3}{8}$ , at which the lower-resolution view becomes dominant in the subjective quality.

Another research studied the subjective impact of uncompressed MR sequences at downsampling ratios of  $\frac{1}{2}$  and  $\frac{1}{4}$  along both coordinate axes [142]. It was found that the perceived sharpness and the subjective image quality of the MR image sequences were nearly transparent at the downsampling ratio of  $\frac{1}{2}$  but dropped slightly at the downsampling ratio of  $\frac{1}{4}$ .

Different qualities between the left and right view can be achieved using different quantization steps too, resulting in different PSNR values for the views. However, it



is not clear what should be the level of this asymmetry. Extensive subjective tests conducted in [125] show that when the reference view is encoded at a sufficiently high quality, encoding the auxiliary view above a low-quality threshold, guarantees that the subjective quality of such asymmetric stereoscopic video can be perceived without a noticeable degradation. It was shown that the low-quality threshold may depend on the 3D display. The subjective results confirm that this threshold should be 21 dB for parallax barrier display and 33 dB for polarized projection display. The authors in [125] further confirmed that at higher bitrates, SNR scaling yields more favorable results compared to spatial resolution reduction. However, the blockiness caused by SNR scaling proved to be more noticeable than blurring caused by downsampling below the mentioned low-PSNR threshold for the auxiliary view. Hence, a conclusion was made that the MR asymmetric coding performs better than asymmetric quality coding at lower bitrates. This is in agreement with previously reported results [20, 143]. This low-quality threshold may depend on the 3D display; e.g. it is about 31 dB for a parallax barrier display and 33 dB for a polarized projection display. Subjective tests conducted in [125] showed that, above this PSNR threshold value, users prefer SNR reduction over spatial resolution reduction on both parallax barrier and polarized projection displays.

## 5.5 Modeling subjective ratings

To reduce the heavy burden of conducting subjective tests, many researchers have considered proposing new algorithms to estimate the subjective quality for both 2D [31, 59, 79, 92, 129, 171, 182] and 3D [10, 12, 17, 22, 53, 54, 60, 111, 126, 130, 132, 133, 168, 187] video content. However, this is still an active and open research topic since no widely accepted and used metric is introduced and agreed between scientists in the research community.

In this thesis, we tried to estimate the subjective quality ratings of MR stereoscopic video presented in [P3], [P10]. A logarithmic relationship between the subjective viewing experience rating and the angular resolution of the lower-resolution view, measured in pixels per degree (PPD) of viewing angle, was observed. As reported in [P10], across all test sequences, high Pearson correlation coefficients were obtained confirming a good estimate of subjective ratings using the logarithmic estimator.

We have also considered using Batch Video Quality Metric (BVQM) software [85] to estimate the subjective scores in this thesis. This software estimates the subjective quality of the input content reporting seven parameters, namely *si\_loss*, *hv\_loss*, *hv\_gain*, *si\_gain*, *chroma\_spread*, *chroma\_extreme*, and *ct\_ati\_gain*. These parameters are shortly described in the following paragraphs while further detailed descriptions of these parameters are presented in [107].

***si\_loss*** detects a decrease or loss of SI e.g. when blurring artifact is introduced.

***hv\_loss*** detects a shift of edges from horizontal and the vertical orientation to

diagonal orientation. For example in the cases where diagonal edges suffer less from blurring effect than horizontal or vertical edges.

*hv\_gain* detects a shift from diagonal edges to the horizontal and vertical ones. This may happen, e.g. when processed video includes tiling or blocking artifacts.

*si\_gain* reports the quality improvement that may result from edge sharpening or enhancements.

*chroma\_spread* detects introduced changes in the way that two-dimensional color samples are spread.

*ct\_ati\_gain* detects severe localized impairments e.g. those produced by digital transmission errors.

*chroma\_extreme* considers the contrast and the temporal information of the input and measures the amount of spatial detail and motion.

In [107], a general video quality metric (VQM) is introduced consisting of a linear combination of these seven parameters. This equation is presented in (5.3). However, it has been shown that *si\_loss*, *hv\_loss*, *hv\_gain*, and *si\_gain* contribute most to the estimation of subjective quality [119]. Hence, a simpler equation considering only these four parameters is introduced in [119] to calculate VQM, as presented in (5.4).

$$\begin{aligned} VQM = & -0.2097 \times si\_loss + 0.5969 \times hv\_loss + 0.2483 \times hv\_gain \\ & + 0.0192 \times chroma\_spread - 2.3416 \times si\_gain \\ & + 0.0431 \times ct\_ati\_gain + 0.0076 \times chroma\_extreme \end{aligned} \quad (5.3)$$

$$\begin{aligned} VQM_{modified} = & -0.2097 \times si\_loss + 0.5969 \times hv\_loss \\ & + 0.2483 \times hv\_gain - 2.3416 \times si\_gain \end{aligned} \quad (5.4)$$

Subjective results reported in [P3] were exploited in this experiment using BVQM software and VQM metric and averaging over both views. The sequences and their associated bitrates used in this experiment are depicted in Table 5.5. Four test cases have been considered in this experiment:

**FR Symmetric:** Full-resolution and same quality for both views

**FR Asymmetric:** Full-resolution for both views and asymmetric quality between views by differing the QP values

Table 5.5: Bitrate selection for different sequences

	Bitrate (Kbps)			
Undo Dancer	1081	1644	2854	5510
Dog	388	598	1078	2000
Newspaper	407	610	1075	2073
Pantomime	1132	1725	3052	5407

Table 5.6: Pearson correlation coefficient between VQM values and mean subjective scores

	FR Symmetric	FR Asymmetric	MR $\frac{1}{2}$	MR $\frac{3}{8}$
Undo Dancer	0.972	0.969	0.939	0.761
Dog	0.928	0.964	0.976	0.834
Newspaper	0.930	0.942	0.965	0.843
Pantomime	0.956	0.939	0.938	0.792

**MR  $\frac{1}{2}$ :** Mixed-resolution where one view been downsampled with ratio  $\frac{1}{2}$  along both directions

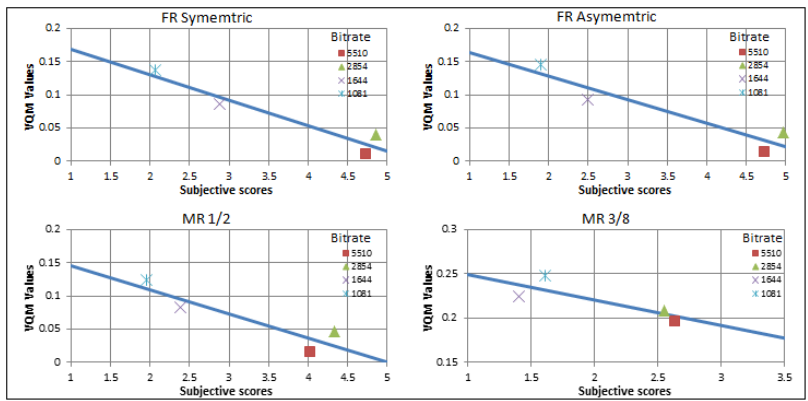
**MR  $\frac{3}{8}$ :** Mixed-resolution where one view has been downsampled with ratio  $\frac{3}{8}$  along both directions

All test cases have been encoded and compared under the same bitrate constraint. VQM is calculated according to the modified VQM (5.4) and the mean value scores have been considered as subjective measures. To ease the similarity of the values, all mean subjective scores have been increased by 3 to become positive and then inversed to have the same characteristics as VQM values. In Figure 5.7 a linear estimate that best fitted the objective and subjective scores has been derived for all sequences. For all test cases except MR  $\frac{3}{8}$ , there is a clear high correlation between the objective and subjective scores. This was further confirmed by calculating the Pearson correlation coefficient for all test cases. These values are reported in Table 5.6. The conclusion reported in [P3] was that the subjective quality of MR  $\frac{1}{2}$  is similar to that of FR cases while MR  $\frac{3}{8}$  underperforms FR schemes. In this experiment, we confirmed that VQM metric can be used for FR symmetric, FR asymmetric, and MR  $\frac{1}{2}$  schemes but is unable to well estimate the subjective experiment scores of MR  $\frac{3}{8}$  scheme which has a clearly lower subjective quality compared to the rest of sequences. Therefore, one can conclude that VQM metric presented in [119] can be utilized to estimate the subjective quality of symmetric and asymmetric FR schemes as well as MR scheme with downsampling ratio of  $\frac{1}{2}$  along each direction.

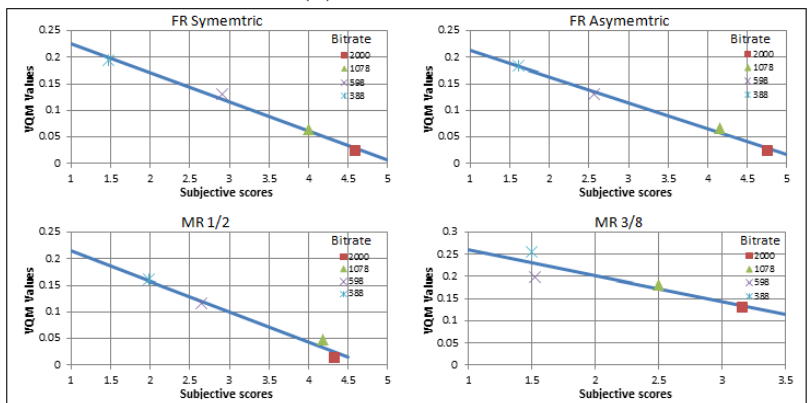
## 5.6 Summary

Considering that several encoding approaches and asymmetric stereoscopic schemes were introduced and/or evaluated subjectively in this thesis, a summary of the main conclusions is provided next.

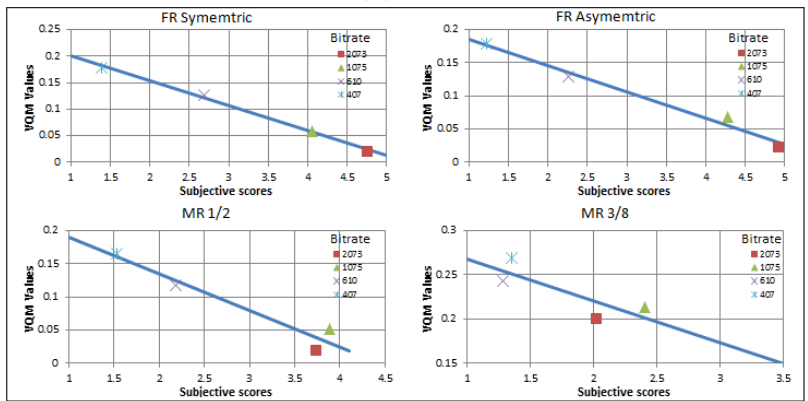
[P6] presents a new MR scheme based on the amount of the SI available in the left and right views of each stereoscopic video. In this scheme, one view is downsampled in the horizontal direction while the other view is downsampled in vertical direction. In this study, each view is evaluated separately and the amount of SI [114] along each direction (vertical and horizontal) is calculated. Comparing these values a



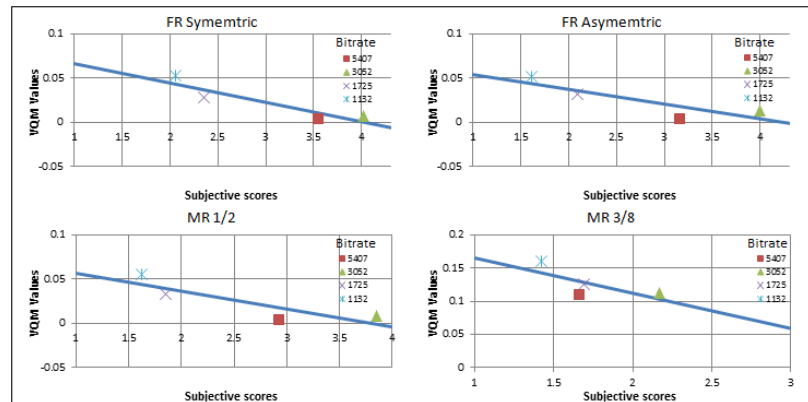
(a) Undo Dancer



(b) Dog



(c) Newspaper



(d) Pantomime

Figure 5.7: Correlation between subjective scores and objective estimates

decision is made for the downsampling direction of each view keeping the maximum accumulated amount of information for the left and right views. Subjective test ratings compare the proposed cross-asymmetric MR scheme with conventional MR (where one view has FR and the other view is downsampled in both directions) and symmetric FR schemes. The results confirm that the proposed method outperforms other schemes and hence, can be considered as a potential MR scheme as it also decreases the number of pixels involved in the encoding and decoding process.

A new asymmetric stereoscopic video coding method is presented in [P9]. This algorithm benefits from two steps: 1) sample domain quantization which in this study is a linear luma value quantization with rounding, and 2) spatial resolution reduction. The quality of the proposed technique was compared subjectively with two other coding techniques: FR symmetric and MR stereoscopic video coding. In most cases (six out of eight) the proposed method achieved a higher mean value for subjective rating compared to the other schemes.

Considering that it has always been required to estimate the subjective quality of videos by an objective metric, in this thesis we considered the results from two sets of experiments presented in [P3] and [P10] and tried to model the subjective ratings with an objective estimate. In this analysis, three downsampling ratios  $\frac{1}{2}$ ,  $\frac{3}{8}$ , and  $\frac{1}{4}$  were used to create the lower resolution view in the asymmetric stereoscopic content. PPD values were calculated and used in the estimation process as they differ for different resolutions. A logarithmic relation was introduced in [P10] to estimate the subjective rating as a function of PPD of a lower resolution view for different sequences and different test setups. The estimated values and actual ratings resulted in high Pearson correlation coefficients, showing that this metric estimates well the subjective ratings under both test conditions and for all test sequences.

# Chapter 6

## Depth-Enhanced Multiview Video Compression

### 6.1 Introduction

The current state of the art multiview video coding standard, MVC [29], is the extension of the H.264/AVC [117]. H.264/AVC and MVC reference softwares [5] were used in some simulations carried out in this thesis. However, conventional frame-compatible stereoscopic video coding techniques, such as the MVC, enable less flexible 3DV displaying at the receiving or playback devices when compared to depth-enhanced MVC. While two texture views, as in the stereoscopic presentation of 3D content, provide a basic 3D perception, it has been discovered that disparity adjustment between views is required for adapting the content to different viewing conditions and also different display types. Moreover, based on personal preferences, it might be desired to have different disparities on the display [138]. Furthermore, ASD technology, as discussed in section 3.4, typically requires the availability of many high-quality views at the decoder/display side prior to displaying. Due to the natural limitations of content production and broadcasting technologies, there is no way that a large number of views can be delivered to the user with the existing video compression standards. In the majority of cases these views are to be rendered in the playback device from the received views. Such needs can be served by coding 3DV data in the MVD format [90, 141] and exploiting the decoded MVD data as input for DIBR [46, 86]. In MVD format, each texture view is accompanied by a respective depth view presenting pixel based associated depth, from which new views can be synthesized using any DIBR algorithm. The encoding process and displaying of depth-enhanced multiview video is presented in Figure 6.1. In the case of stereoscopic presentation, the desired views for the selected disparity and hence the depth perception will be chosen from the decoded and synthesized views at the display side. Moreover, considering the ASD presentation, based on the required number of views, a subset of the total decoded and synthesized views will be utilized.

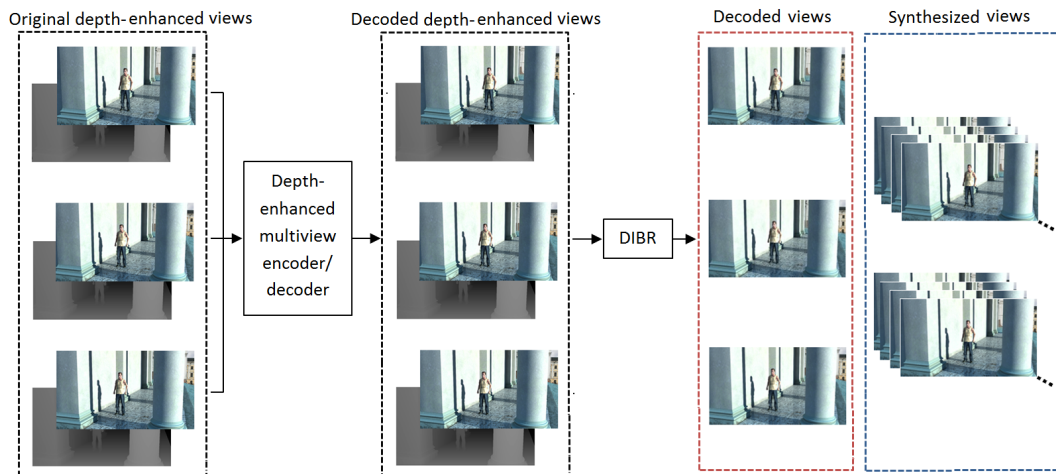


Figure 6.1: Encoding and synthesis process for a depth-enhanced multiview video

Considering that MVC was not targeting depth-enhanced multiview format, it is not optimized to encode both texture and depth maps. As a result, there have been standardization efforts towards depth-enhanced video coding and MPEG issued a Call for Proposals (CfP) for 3DV coding technology in March 2011 [3]. The target of this CfP was to satisfy the following two ideas: (1) enabling a variety of 3D applications and display types including a varying baseline to adjust the depth perception, (2) supporting multiview ASDs.

Two projects covered by CfP are described in the following paragraphs.

The CfP invited submissions in two categories, the first is compatible with H.264/AVC and the second is compatible with the High Efficiency Video Coding (HEVC) [146] standard. A depth-enhanced extension for MVC, abbreviated MVC+D, specifies the encapsulation of MVC-coded texture and depth views into a single bitstream [30, 149]. The utilized coding technology is identical to MVC, and hence MVC+D is backward-compatible with MVC and the texture views of MVC+D bitstreams can be decoded with an MVC decoder. The MVC+D specification was finalized technically in January 2013. The reference software [5] implementation of MVC+D has been used in several simulations in this thesis [P1], [P4], and [P5].

Joint Collaborative Team on 3D Video (JCT-3V) is an organization targeting ongoing video coding development extension of H.264/AVC, referred here to as 3D-AVC. This development exploits redundancies between texture and depth and includes several coding tools that provide a compression improvement over MVC+D. The specification requires that the base texture view is compatible with H.264/AVC and compatibility of dependent texture views to MVC may optionally be provided. 3D-AVC is planned to be finalized technically in November 2013. The reference software implementation of H.264/AVC has been used in few publications in this thesis [P7] and [P8].

## 6.2 Depth map

A depth map presents the values related to the distance of the surfaces of the scene objects from the view point of the recording camera or observer. Depth maps are usually presented with gray scale images (8 bits per pixel) where closer objects to the camera are represented with larger values and objects farthest away from the camera are represented with the smallest value, i.e. they appear as black (0 gray level) pixels in the depth image. Another approach to represent the depth values of different views in the stereoscopic or multiview case is to report the disparity between pixels of each view to the adjacent view instead of the actual depth values. The following equation (6.1) shows how depth values are converted to disparity:

$$D = f \times l \times \left( \frac{d}{2^N - 1} \times \left( \frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right) + \frac{1}{Z_{far}} \right) \quad (6.1)$$

where:

$D$  = disparity value

$f$  = focal length of capturing camera

$l$  = translational difference between cameras

$d$  = depth map value

$N$  = number of bits representing the depth map values

$Z_{near}$  and  $Z_{far}$  are the respective distances of the closest and farthest objects in the scene to the camera (mostly available from the content provider), respectively.

In most of the experiments carried out in the thesis, depth maps are considered unless the use of disparity is explicitly described. Depth maps can be considered approximately piecewise planar surfaces consisting of highly homogeneous regions separated by strong contours. As a result, one can conclude that preserving better the contours increases the usefulness of depth maps in virtual view synthesis. This is due to the fact that small miss-adjustments in an area having a similar depth might not have an annoying effect as it could have along a strong contour separating the foreground and the background of a picture. This can be confirmed while observing the ongoing research on both segmentation based compression methods applied to depth maps [67, 99, 121] and edge adaptive algorithms that target to preserve edges as accurately as possible [135, 136].

A number of approaches have been proposed for representing depth picture sequences, including the use of auxiliary depth map video streams, MVD [90], and layered depth video (LDV) [131], which are described briefly in the sequel. The depth map video stream for a single view can be regarded as a regular monochromatic video stream and is coded with any video codec. The essential characteristics of the depth map stream, such as the minimum and maximum depth in world coordinates, can be indicated in messages formatted according to the MPEG-C Part 3 standard [18]. In the MVD representation, the depth picture sequence for each texture view is coded with any video codec, such as MVC. In the LDV representation, the texture and depth of the central view are coded conventionally, while the



texture and depth of the other views are partially represented and cover only the dis-occluded areas required for correct view synthesis of intermediate views.

### 6.3 Synthesizing virtual views

The term view synthesis (or view rendering) refers to the generation of a new view based on one or more existing or received views. Although differing in details, most of the view synthesis algorithms utilize 3D warping based on explicit geometry, i.e., depth images. Typically in depth images each texture pixel is associated with a depth pixel indicating the distance or the z-value from the camera to the physical object from which the texture pixel was sampled. Different basic algorithms for view synthesis are proposed. McMillan's approach [86] uses a non-Euclidean formulation of the 3D warping, which is efficient under the condition that the camera parameters are unknown or the camera calibration is poor. Mark's approach [83], however, strictly follows the Euclidean formulation, assuming that the camera parameters for the acquisition and view interpolation are known. For virtual view rendering, one pair of neighboring original camera views and their associated depth maps are used to render arbitrary virtual views on a specified camera path between them. Depth images are used to assist in correct synthesis of the virtual views.

The relation between points in a 3D scene space and the values available in a depth map are defined by the projection matrix. In the first step, for each camera the depth maps are unprojected, resulting in a colored 3D particle cloud. Then, in the location of each virtual camera a projection matrix is calculated from the two projection matrices of cameras by spherical linear interpolation (SLERP) [140] and linear interpolation (LERP) [35]. Using the projection matrix of virtual cameras we have the ability to render the virtual view weighting according to the position of the virtual camera. Figure 6.2 shows the high level scheme of view synthesis.

Occlusions, pinholes, and reconstruction errors are the most common artifacts introduced in the 3D warping process [83]. These artifacts occur more frequently in the object edges, where pixels with different depth levels may be mapped to the same pixel location in the virtual image. When those pixels are averaged to reconstruct the final pixel value at the pixel location in the virtual image, an artifact might be generated, since pixels with different depth levels usually belong to different objects. There exist several different techniques to perform view synthesis and filling holes and disoccluded areas e.g. [68, 100, 155, 164, 191].

In all conventional schemes each texture view is accompanied with a depth view which is used in the view synthesis process. However, in publication [P4] it has been shown that it is redundant to have the same number of depth views as texture views and the number of depth views can be reduced without sacrificing the quality of stereoscopic views. The simulation results confirm the efficiency of the proposed format reporting 1% to 7% of Bjontegaard delta bitrate reduction [14] for the baseline, and disparity adjustments from 50% to 100% of the coded baseline.

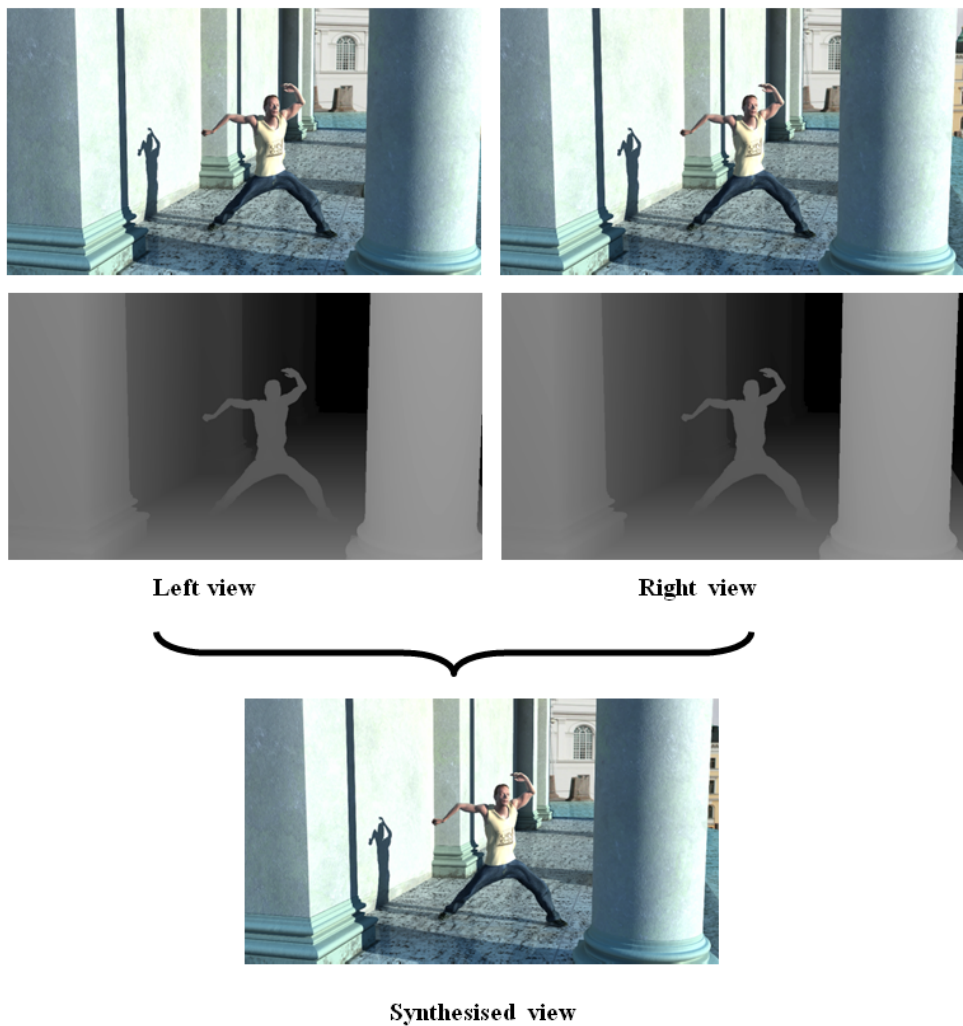


Figure 6.2: A synthesized view

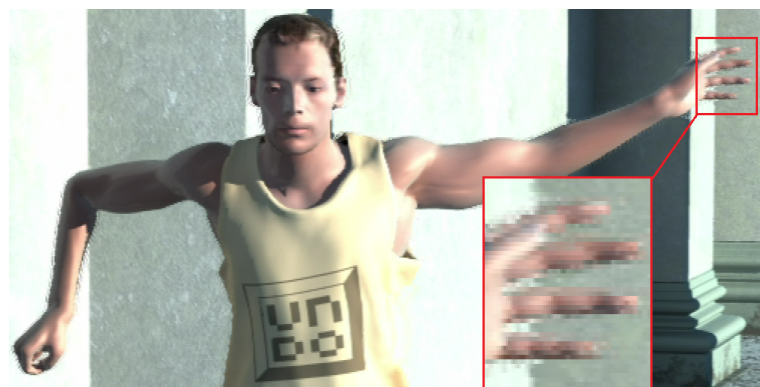
## 6.4 Quality dependency of rendered views

The quality of rendered views depends on both texture views and depth maps as well as the rendering algorithm used. Clearly more accurate texture views compared to the original views provide rendered views with a higher quality. However, this is not always the case for depth maps. A visualization of this concept is presented in Figure 6.3 where the same texture along with different depth maps for rendering is used. For synthesizing these views the VSRS ver 3.5 [154] is used. In Figure 6.3a, the original depth map is used, while in Figure 6.3b, the depth map was low-pass filtered to smooth the edges prior to rendering. Comparing the subjective quality of the two frames, one concludes that depth maps with sharper edges produce a higher quality rendered views around edges, while areas where no depth change occurs remain unchanged.

To provide a general view about the quality dependency of synthesized views on the amount of spatial resolution of texture and depth maps, two sets of experiments were conducted in this thesis. For a depth-enhanced multiview test set including 7 sequences (as specified by MPEG Common Test Conditions (CTC) [6]) and three views (reference views) per sequence, view synthesis using VSRS ver 3.5 [154] between the reference views was performed creating three equally spaced virtual views for each couple of reference views resulting into six virtual views in total. No compression was applied to the texture or depth views while in one set of experiments, the spatial resolution of texture views was reduced to half along each direction (QR Texture) and in the other case the same spatial resolution reduction was applied to only the depth views (QR Depth). In both cases the PSNR of the synthesized views was calculated against the same views synthesized from the original FR texture and the depth views. The average PSNR over 6 synthesized views is reported in Table 6.1. The result of these experiments shows that while texture spatial resolution drastically affects the PSNR of synthesized views, the spatial resolution of the depth



(a)



(b)

Figure 6.3: Rendered view from (a) original depth map and (b) low-pass filtered depth map

Table 6.1: PSNR of synthesized views based on spatial resolution of reference texture and depth views

Sequence	PSNR (dB)	
	QR Texture	QR Depth
Poznan Hall2	44.18	52.30
Poznan Street	39.14	47.99
Undo Dancer	32.55	38.38
Ghost Town Fly	35.47	43.79
Kendo	45.77	49.79
Balloons	44.58	48.49
Newspaper	40.15	42.93

views has also a considerable effect on the quality of the synthesized views.

## 6.5 Depth map compression

In different 3DV scenarios, depth maps are used as supplementary data along with texture views, for synthesizing new images but not to be directly observed by the users. Thus, in depth map compression, the goal is to maximize the perceived visual quality of the synthesized views rather than improving the visual quality of the depth maps themselves [89]. Traditional video coding methods have been designed to operate through a Rate-Distortion Optimization (RDO) of coded data and a pixel-based distortion introduced by the codec, e.g. Sum of Absolute Differences (SAD) or Mean Square Error (MSE). However, it has been shown that coding distortions introduced to depth maps typically have a non-linear impact on the visual quality of the synthesized views [69]. Therefore, depth map compression using traditional RDO algorithms might result in suboptimal performance of 3DV coding systems [69]. In the following sub-sections, two traditional and commonly used pre- and post-processing methods applied to depth maps are presented.

As described in section 6.4, the quality of the synthesized views depends on the quality of the depth map coding and the quality of the coding applied to the original color view used as a reference. As for depth maps, errors in the depth map close to a sharp edge, having considerable different depth values on the sides, can result in severe rendering artifacts, while errors on a smooth and homogenous area may have negligible subjective influence on the quality of the synthesized view. These ideas are exploited in several active research studies where either an encoding algorithm based on depth map segmentation is introduced [67, 99, 121] or an edge adaptive algorithm is proposed [135, 136]. Considering the structure of depth maps, it is wise to spend more bits to encode the edges and try to preserve them as much as possible while sacrificing the relative quality of the homogeneous areas. This claim is well

illustrated in Figure 6.3.

### 6.5.1 Depth map filtering

Considering the nature of depth maps as described in section 6.2, it has been shown that smoothing the depth maps prior to utilization in DIBR algorithms for synthesizing virtual views can increase the subjective quality [151]. Depth information may be generated using a specific depth sensor [76, 190] or it can be generated in a per-pixel depth estimation process based on texture views. In both cases such filtering will decrease the existing noise and artifacts which can result in poorer quality of synthesized views. Moreover, authors in [151] conducted a series of subjective tests concluding that increasing levels of smoothing applied to depth maps increased the perceived image quality scores.

### 6.5.2 Depth down/up sampling

Traditional image downsampling techniques use linear filters, whereas depth downsampling should preserve sharp edges of the depth data. Hence, edge-preserving downsampling for depth has been considered. For example, in [157, 176], the median value of a  $N \times N$  window was chosen as the most representative value to be used at reduced-resolution depth map (where the factor  $N$  specifies the downsampling ratio along each direction).

Similarly to downsampling, also upsampling should preserve depth edges. In various works, e.g. [118, 176], cross-component bi-lateral filtering has been used for depth upsampling or filtering. In a cross-component bi-lateral filter, the similarity of co-located texture samples is used to derive filter weights for depth in addition to the conventional filtering window applied spatially for the depth samples.

Another approach for depth upsampling was used in [157]. For the reconstruction process, the decoded depth data is first up-sampled using a nearest neighbor filter, which is followed by post-processing using a median filter, a frequent-low-high filter and a cross-component bi-lateral filter. The 2D median filter is used to smooth blocking artifacts caused by depth down-sampling. The frequent-low-high filter is a non-linear filter used to recover object boundaries, which results into selecting either the most frequently occurring sample value below or above the median sample value within a filter window. More information on the frequent-low-high filter is available in [101]. The bilateral filter is used to eliminate the errors still present after both filtering procedures. In [157], all post-processing filters were applied with a  $7 \times 7$  window size.

In [94], a cross-trilateral filter was used as a post-processing step for depth estimation to improve the quality of the estimated depth maps. The proposed filter adds a depth based weight to a conventional cross-component bi-lateral filtering approach. The depth based weight depends on both the depth similarity and the confidence of the depth value correctness, where the confidence is derived from the

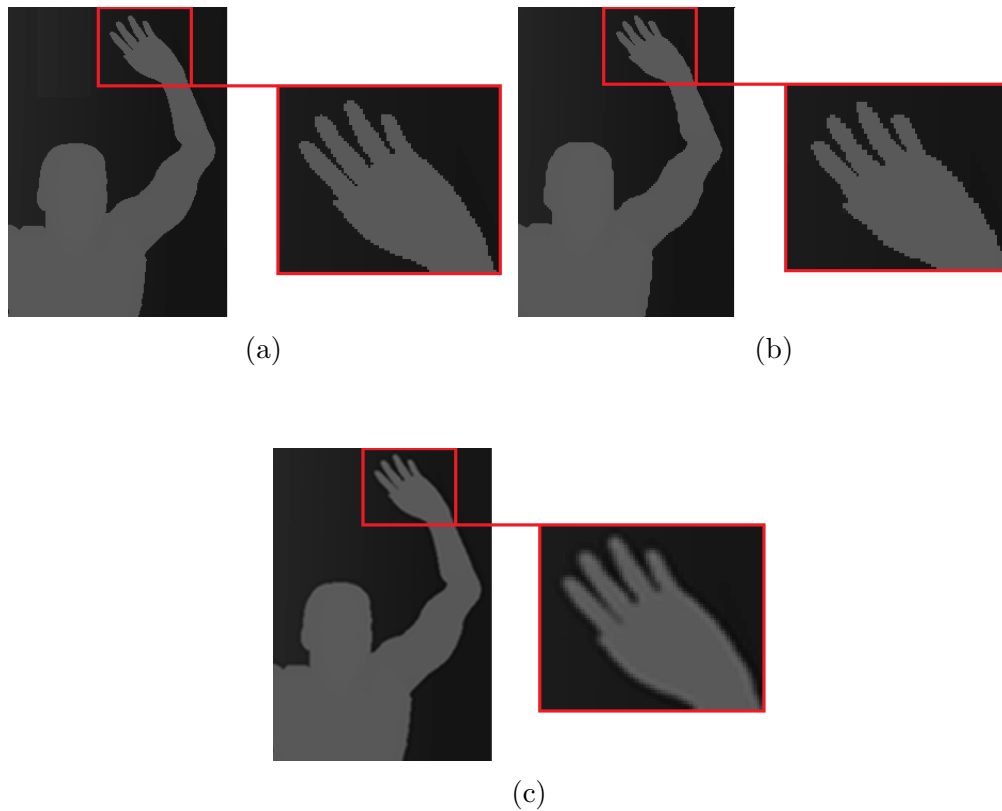


Figure 6.4: Resampled depth maps (a) original, (b) proposed method in [P1], (c) JSVM

correlation of the disparity/depth estimations obtained from left-to-right and right-to-left correspondences.

In this thesis a novel non-linear resampling approach for depth maps is presented [P1]. Figure 6.4 depicts the performance of the proposed non-linear depth map down/up sampling introduced in [P1] by showing the original, resampled depth maps using the resampling in the reference JSVM software [1], and the resampled depth maps using the proposed method. In both resampling cases (Figures 6.4b and 6.4c), the original depth map has been down sampled with ratio  $\frac{1}{2}$  along each direction and then upsampeld to FR. This figure presents an example of depth map edge preserving during the resampling process since as illustrated in Figure 6.3, the quality of synthesized views, using resampled depth maps with more accurate edges, is considerably higher compared to the quality of synthesized views rendered from depth maps having smooth edges.

## 6.6 Using asymmetry in multiview video compression

In section 5.3, we explained the advantages of exploiting asymmetry in stereoscopic video compression. The same idea can be applied to multiview video as well. Several cases of asymmetric multiview video compression are described in the following sub-sections.

### 6.6.1 Asymmetric quality

In a three-view video format, asymmetry between views can be easily achieved by applying coarser quantization steps to some views compared to the other views. This will result in a bitrate reduction in the asymmetric views and hence a decrease in the total bitrate required to encode the 3D video. In publication [P8], the two side-views of the three-view format of a depth-enhanced multiview video were encoded with a higher QP compared to the central view. Furthermore, several views were synthesized in between the coded views and among them a stereopair was selected having a suitable baseline for conventional stereoscopic displays. A subjective comparison was conducted between the stereopair from the proposed scheme and the stereopair from the scheme where all views were encoded using the same QP as that of the central view in the proposed scheme. The results confirmed that in average 20% bitrate reduction can be achieved by exploiting such an asymmetric scheme. Furthermore, in [P8] the usability of such scheme for ASD utilization was objectively confirmed.

### 6.6.2 Mixed-resolution texture

As explained in detail in section 5.3, asymmetric quality with different presentations and implementations can be exploited to increase the coding efficiency in the compression of stereoscopic video content. However, in this sub-section, the same idea is deployed in a depth-enhanced multiview scheme, where more than two views are involved, e.g., a test scenario where three texture views and three associated depth maps are encoded and several virtual views are rendered in between the coded views. In this case the quality of all coded views and rendered views is of importance depending on the application in which the codec is used. Namely, if a stereoscopic display is targeted and the baseline separation is similar to those between the coded views, there is no need to render any virtual views. However, if the target is ASD or a baseline adjustable stereoscopic content, the rendering process is inevitable.

In publication [P5], a test scheme where the resolution of the side views is reduced to half along each direction is introduced. Therefore, both side views have quarter spatial resolution compared to the central view. Moreover, two inter-view prediction schemes are introduced and the results confirm that on average 4% and 15% delta bitrate reduction are achieved.

## 6.7 Video compression artifacts

There exist several different coding artifacts due to compression applied to the video content [188]. In general these artifacts can be divided to two categories:

- 1 Spatial artifacts e.g. blocking, blurring, and ringing.
- 2 Temporal artifacts e.g. color bleeding, temporal artifacts, false edges, motion jerkiness, mosquito noise, flickering, and bumping.

In this thesis the distortions introduced to the content based on the blocking and blurring artifacts are mainly considered since variable quantization steps and downsampling ratios are applied for further compression of the video sequences. There have been several post processing algorithms proposed to remove blocking artifacts in the spatial domain [24, 40, 49, 77, 81, 185, 189] or in the transformed domains e.g., discrete cosine domain (DCT) domain [28, 82] or wavelet domain [63, 78, 184]. Compared to extensive research to remove blocking artifact as a post processing technique, the de-blurring process is mainly focused on optimized downsampling and upsampling filters [48, 137, 139]. The blurring is mostly introduced to images because of low-pass filtering or alternatively down sampling prior to encoding and upsampling after decoding. In the video compression field and especially when subjective quality assessment is required, knowledge about these artifacts is vital as they do not necessarily provide similar subjective and objective quality degradations. Figure 6.5 depicts blocking and blurring artifact to the same input. The subjective preference and hence, compromise between these two artifacts has been studied in publications [P3], [P6], [P7], and [P9].

## 6.8 Summary of subjectively assessed experiments

Considering that several encoding approaches and asymmetric stereoscopic schemes were introduced and/or evaluated subjectively in this thesis, the main findings are summarized next.

In [P7], the aim was to develop a proper technique to decide which downsampling ratio should be applied to texture views prior to encoding to increase the subjective quality of the decoded videos under the same bitrate constraint. Two methods are presented in [P7]: 1) an MSE based technique, and 2) a frequency based technique. Considering the results of the conducted subjective tests on different resolutions, we found out that the MSE based metric is weakly correlated with the resolution selection based on the subjective quality. However, the frequency-based distortion metric was able to well estimate the selection of downsampling ratios in agreement with the selection based on the subjective test. Hence, the proposed method can be considered as a potential candidate metric to assure the best perceived quality by a proper selection of the texture view resolution prior to encoding.



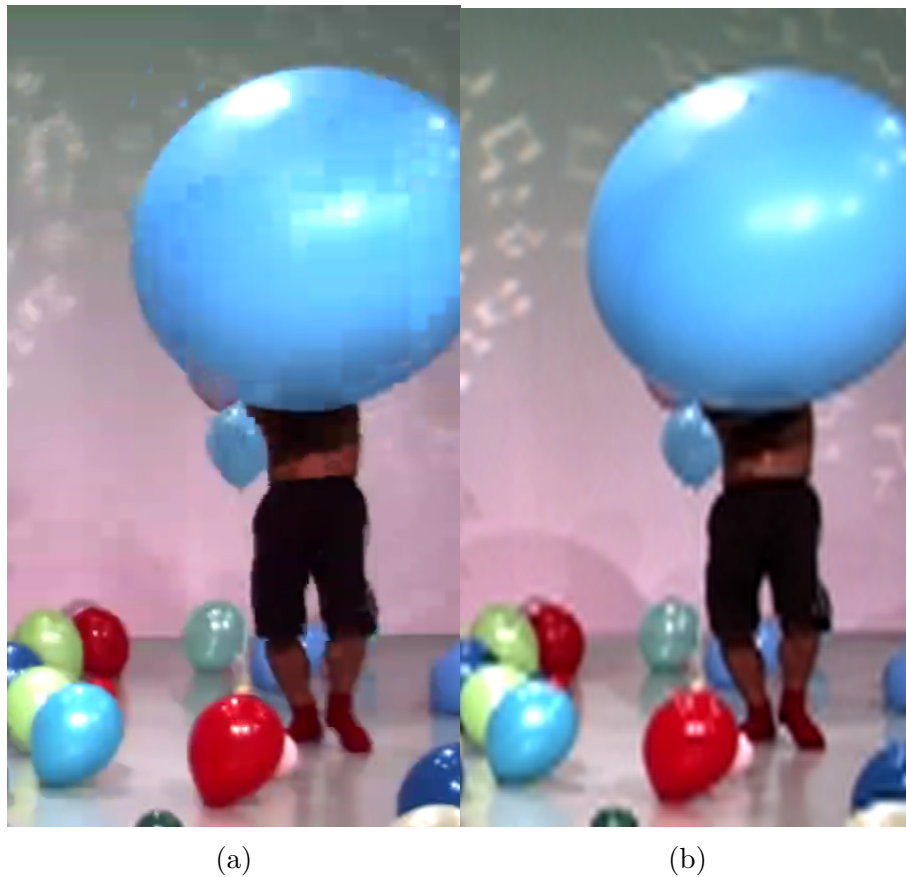


Figure 6.5: Encoding artifacts(a) blocking and (b) blurring

Random noise in captured multiview content is a source of inconsistency between views. In [122] a locally adaptive filtering in 3D DCT domain was introduced and utilized in the pre-processing stage to improve the encoding performance. In [9], we applied the de-noising algorithm introduced in [122] to a 3-view multiview test scenario comparing the perceived subjective quality of synthesized views from those three views with non-denoised original synthesized views. A set of subjective tests confirmed that up to 11.7% average bitrate reduction can be achieved without any noticeable subjective quality degradation. Hence, it was subjectively assured that applying the proposed de-noising algorithm prior to encoding is capable of decreasing the required bitrate for coding the same content while negligible reduction in subjective quality is introduced.

In [P8], we presented an asymmetric quality three-view scenario for depth-enhanced multiview targeting a lower bitrate under the same subjective quality constraint. In this study, out of three texture views, the side views were coded with coarser quantization steps and hence, had lower quality compared to the central view. Following this, taking into account a suitable baseline for conventional 3D displays, a suitable stereopair was selected from synthesized views between refer-

---

ence views and the perceived quality of this stereopair was subjectively assessed. The results confirmed that on average 20% bitrate reduction can be achieved with a negligible penalty on the subjective quality of the sequences. Moreover, we objectively evaluated the performance of the same asymmetric scheme to be used on ASDs confirming that it is also beneficial when the same content is targeting ASDs compared to the case where the symmetric encoding scheme is used.



# Chapter 7

## Conclusion and Future Work

In this thesis 3DV compression was tackled considering different formats, namely, stereoscopic video and depth-enhanced multiview video. In general the methods utilized for stereoscopic video compression can be extended to multiview video compression. Moreover, higher efficiency in depth-enhanced multiview video compression will result in better performance for multiview ASD too. Hence, Stereoscopic video compression can be considered as a basis of 3D content compression while targeting different applications and a broad variety of display devices.

The research carried out in this thesis introduced novel compression techniques for 3D content and investigated several related topics, e.g., estimation of the subjective scoring with objective calculations, simultaneous presentation of the same content for both 2D and 3D perception, and depth map resampling targeting higher quality for synthesized views. All schemes introduced in the thesis achieved a higher performance compared to the conventional reference under the same criteria e.g. obtaining better subjective quality or less bitrate under equal bitrate or the same subjective quality constraint, respectively.

A large part of the thesis focused on exploring different types of asymmetry in 3D video compression. A number of different asymmetric schemes for 3DV compression were introduced and evaluated. The conclusions for all methods confirmed that asymmetric video compression is a promising technique, where the required bitrate or the coding complexity was reduced. Since no standardized or widely used objective metric for evaluating the perceived quality of asymmetric quality 3D content is known to the research community, in this thesis several formal and systematic subjective test experiments were conducted to evaluate the quality of the codec being tested.

Finally, the combination of objective and subjective assessments reported in this thesis confirmed that the proposed algorithms are superior to conventional approaches from bitrate reduction and complexity points of view. Moreover, new schemes and formats for presentation of 3D content have been introduced and evaluated, targeting for specific applications.

---

## 7.1 Future work

Future work includes deeper studies on the HVS and the reaction of its fusion system to different quality changes introduced between stereoscopic views. Moreover, the proposed methods and algorithms in this thesis can be extended considering different tuning parameters than those already used. This will evaluate the robustness of the proposed methods and potentially enables the introduction of higher performance schemes.

The different types of asymmetry introduced in this thesis can be extended by introducing and evaluating new schemes as well as exploiting different combinations of the schemes presented in this thesis (e.g. sample value quantization and LPF or Chroma sampling and LPF).

Considering the large amount of subjective quality assessments conducted in this thesis, a proper database is available to authors for further research and analysis investigating different available objective metrics. Moreover, since all test material and details of the test setup are known, it is possible to produce new objective metrics targeting accurate estimation of available subjective scores and to verify their validity under different conditions.

Furthermore, the subjective results reported in this thesis should be confirmed under different test setups (e.g. viewing distance, display resolution, viewing conditions, and/or test duration), alternative view synthesis algorithms, or test material (e.g. different bitrates and varied content by duration, resolution, or frame rate).

Considering the ever increasing demand to view 3D content without glasses, an important future continuation of this thesis is to test the usability of the proposed technique for simultaneous 2D and 3D visualization of stereoscopic content when used in ASD. This includes depth-enhanced compression techniques introduced in this thesis as they are able to feed the ASD with arbitrary required number of views at the decoder side. This thesis lacks subjective quality evaluation performed with ASD and hence, the conclusions obtained with objective metrics, can be further confirmed by conducting subjective tests on the same content.

# Bibliography

- [1] *JM Reference Software (Last checked: May,2013)*. [Online]. Available: <http://iphone.hhi.de/suehring/tml/download>.
- [2] *Coding of Still Pictures (Last checked: May,2013)*. ISO/IEC JTC 1/SC 29/WG 1, 2005. [Online]. Available: <http://www.itscj.ipsj.or.jp/sc29/29w12901.htm>.
- [3] *Call for proposals on 3D video coding technology*. Moving Picture Experts Group - document N12036, 2011. [Online]. Available: [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3dv-cfp.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3dv-cfp.zip).
- [4] *Sky 3D pub*. British Sky Broadcasting Group plc and Sky IP International Limited, 2012. [Online]. Available: <http://3d.sky.com/pubfinder/>.
- [5] “Test model for avc based 3d video coding,” in *ISO/IEC JTC1/SC29/WG11 MPEG2012/N12558*, 2012.
- [6] “Common test conditions for 3dv experimentation,” *ISO/IEC JTC1/SC29/WG11 MPEG2012/N12560*, February, 2012.
- [7] P. Aflaki, M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, “Subjective study on compressed asymmetric stereoscopic video,” in *17th IEEE International Conference on Image Processing (ICIP)*. IEEE, Hong Kong, September, 2010, pp. 4021–4024.
- [8] —, “Impact of downsampling ratio in mixed-resolution stereoscopic video,” in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, Tampere, Finland, June, 2010, pp. 1–4.
- [9] P. Aflaki, D. Rusanovskyy, T. Utriainen, E. Pesonen, M. Hannuksela, S. Jumisko-Pyykko, and M. Gabbouj, “Texture denoising utilized in depth-enhanced multiview video coding,” in *Picture Coding Symposium (PCS)*. IEEE, Krakow, Poland, May, 2012, pp. 85–88.
- [10] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, “No-reference stereoscopic image quality assessment,” in *IST/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 75 240T–75 244T.

- 
- [11] A. Aksay, C. Bilen, and G. B. Akar, "Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images," in *Proc. 13th European Signal Processing Conference, EUSIPCO*, 2005, Antalya, Turkey.
- [12] B. Alexandre, L. Patrick, C. Patrizio, and C. Romain, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, 2009.
- [13] T. ōOkoshi, *Three-dimensional imaging techniques*. Academic Press (New York), 1976.
- [14] G. Bjontegard, "Calculation of average psnr differences between rd-curves," *ITU-T VCEG-M33*, 2001.
- [15] R. Blake, "Threshold conditions for binocular rivalry." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 2, pp. 251–257, 1977.
- [16] —, "A neural theory of binocular rivalry." *Psychological review*, vol. 96, no. 1, p. 145, 1989.
- [17] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," in *IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE, Denver, Colorado, March, 2006, pp. 218–222.
- [18] A. Bourge, J. Gobert, and F. Bruls, "Mpeg-c part 3: Enabling the introduction of video plus depth contents," in *Proc. of IEEE Workshop on Content Generation and Coding for 3D-television, Eindhoven, The Netherlands*, 2006.
- [19] A. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1132–1144, 2003.
- [20] H. Brust, A. Smolic, K. Mueller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, May, pp. 1–4.
- [21] J. Bullinaria *et al.*, "Learning and evolution of control systems," *Neural network world*, vol. 10, no. 4, pp. 535–544, 2000.
- [22] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proceedings of 15th European Signal Processing Conference (EUSIPCO)*, Poznan, Poland, September, 2007.
- [23] R. H. Carpenter, "Movements of the eyes (2nd revision)." Pion Limited, 1988.

- 
- [24] R. Castagno, S. Marsi, and G. Ramponi, "A simple algorithm for the reduction of blocking artifacts in images and its implementation," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 3, pp. 1062–1070, 1998.
- [25] O. Castle, "Synthetic image generation for a multiple-view autostereo display," Ph.D. dissertation, University of Cambridge, 1995.
- [26] D. Chandler, *Visual Perception (Introductory notes for media theory students) (Last checked: May, 2013)*. University of Wales, Aberystwyth, 2008. [Online]. Available: <http://www.aber.ac.uk>.
- [27] D. Chandler and S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.
- [28] T. Chen, H. Wu, and B. Qiu, "Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 5, pp. 594–602, 2001.
- [29] Y. Chen, Y. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging mvc standard for 3d video services," *EURASIP Journal on Applied Signal Processing*, vol. 2009, pp. 1–8, no. 1, January, 2009, article ID 786015.
- [30] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori, "Overview of the mvc+d 3d video coding standard," *to appear in Journal of Visual Communication and Image Representation*, 2013.
- [31] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.
- [32] A. Çöltekin, *Foveation for 3D visualization and stereo imaging*. Helsinki University of Technology, 2006.
- [33] R. L. Cook, T. Porter, and L. Carpenter, "Distributed ray tracing," in *ACM SIGGRAPH Computer Graphics*, vol. 18, no. 3. ACM, 1984, pp. 137–145.
- [34] H. Coolican, *Research methods and statistics in psychology*. Hodder Education, 2009.
- [35] E. B. Dam, M. Koch, and M. Lillholm, *Quaternions, interpolation and animation*. Datalogisk Institut, Københavns Universitet, 1998.
- [36] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, 2000.



- 
- [37] D. J. DeBitetto, “Holographic panoramic stereograms synthesized from white light recordings,” *Applied Optics*, vol. 8, no. 8, pp. 1740–1741, 1969.
- [38] N. Dodgson, “Analysis of the viewing zone of the cambridge autostereoscopic display,” *Applied Optics*, vol. 35, no. 10, pp. 1705–1710, 1996.
- [39] —, “Autostereoscopic 3d displays,” *Computer*, vol. 38, no. 8, pp. 31–36, 2005.
- [40] I. Draft, “Recommendation h. 263, video coding for low bit rate communication,” *International Telecommunication Union*, 1995.
- [41] M. Eckert and A. Bradley, “Perceptual quality metrics applied to still image compression,” *Signal Processing*, vol. 70, no. 3, pp. 177–200, 1998.
- [42] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, “New full-reference quality metrics based on hvs,” in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, January, 2006.
- [43] E. Ekmekcioglu, S. Worrall, and A. Kondoz, “Bit-rate adaptive downsampling for the coding of multi-view video with depth information,” in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2008*. IEEE, May, 2008, pp. 137–140.
- [44] A. Eskicioglu and P. Fisher, “Image quality measures and their performance,” *IEEE Transactions on Communications*, vol. 43, no. 12, pp. 2959–2965, 1995.
- [45] D. Ezra, G. Woodgate, B. Omar, N. Holliman, J. Harrold, and L. Shapiro, “New autostereoscopic display system,” in *IS&T/SPIE’s Symposium on Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1995, pp. 31–40.
- [46] C. Fehn, “Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv,” in *Electronic Imaging*. International Society for Optics and Photonics, May, 2004, pp. 93–104.
- [47] J. Ferwerda, “Elements of early vision for computer graphics,” *Computer Graphics and Applications, IEEE*, vol. 21, no. 5, pp. 22–33, 2001.
- [48] T. Frajka and K. Zeger, “Downsampling dependent upsampling of images,” *Signal Processing: Image Communication*, vol. 19, no. 3, pp. 257–265, 2004.
- [49] X. Gan, A. Liew, and H. Yan, “Blocking artifact reduction in compressed images based on edge-adaptive quadrangle meshes,” *Journal of Visual Communication and Image Representation*, vol. 14, no. 4, pp. 492–507, 2003.

- 
- [50] B. Girod, “What’s wrong with mean-squared error?” in *Digital images and human vision*. MIT press, 1993, pp. 207–220.
- [51] E. Goldstein, “Sensation and perception (rev. ed.),” *Pacific Grove, CA: Wadsworth-Thomson Learning*, 2002.
- [52] V. Gopinathan, Y. P. Tsividis, K.-S. Tan, and R. K. Hester, “Design considerations for high-frequency continuous-time filters and implementation of an antialiasing filter for digital video,” *IEEE Journal of Solid-State Circuits*, vol. 25, no. 6, pp. 1368–1378, 1990.
- [53] P. Gorley and N. Holliman, “Stereoscopic image quality metrics and compression,” *Stereoscopic Displays and Applications XIX*, pp. 680 301–680 305, San Jose, CA, USA, January, 2008.
- [54] K. Ha and M. Kim, “A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video,” in *18th IEEE International Conference on Image Processing*. IEEE, September, 2011, pp. 2525–2528.
- [55] Y. HaCohen, R. Fattal, and D. Lischinski, “Image upsampling via texture hallucination,” in *IEEE International Conference on Computational Photography (ICCP)*. IEEE, Cambridge, March, 2010, pp. 1–8.
- [56] P. Hanhart, F. De Simone, and T. Ebrahimi, “Quality assessment of asymmetric stereo pair formed from decoded and synthesized views,” in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, Yarra Valley, Australia, July, 2012, pp. 236–241.
- [57] P. Hanhart and T. Ebrahimi, “Quality assessment of a stereo pair formed from decoded and synthesized views using objective metrics,” in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, Tampere, Finland, May, 2010, pp. 1–4.
- [58] H. Helmholtz, *Treatise on physiological optics*. 1867, edition reprinted Thoemmes Press, 2000.
- [59] S. Hemami and A. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal processing: Image communication*, vol. 25, no. 7, pp. 469–481, 2010.
- [60] C. Hewage, S. Worrall, S. Dogan, and A. Kondoz, “Prediction of stereoscopic video quality using objective quality models of 2-d video,” *Electronics letters*, vol. 44, no. 16, pp. 963–965, 2008.
- [61] I. Howard, *Seeing in depth, Vol. 1: Basic mechanisms*. University of Toronto Press, 2002.

- 
- [62] I. Howard and B. Rogers, *Binocular vision and stereopsis*. Oxford University Press, USA, 1995.
- [63] T. Hsung, D. Pak-Kong Lun, and W. Siu, "A deblocking technique for block-transform compressed image using wavelet transform modulus maxima," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1488–1496, 1998.
- [64] D. H. Hubel, *Eye, brain, and vision*. Scientific American Library/Scientific American Books, 1995.
- [65] W. IJsselsteijn, P. Seuntiëns, and L. Meesters, "Human factors of 3d displays," *3D Videocommunication: Algorithms, Concepts, and Real-Time Systems in Human-Centred Communication*, pp. 219–234, 2005.
- [66] W. IJzerman, S. Zwart, and T. Dekker, "7.4: Design of 2d/3d switchable displays," in *SID Symposium Digest of Technical Papers*, vol. 36, no. 1. Wiley Online Library, 2005, pp. 98–101.
- [67] F. Jager, "Contour-based segmentation and coding for depth map compression," in *Visual Communications and Image Processing (VCIP)*. IEEE, November, 2011, pp. 1–4.
- [68] M. Karsten, S. Aljoscha, D. Kristina, M. Philipp, K. Peter, W. Thomas, *et al.*, "View synthesis for advanced 3d video systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2009.
- [69] W. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," *SPIE Visual Information Processing and Communication*, vol. 7543, pp. 75 430B,75 430B–10, 2010.
- [70] J. Konrad and M. Halle, "3-d displays and signal processing," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 97–111, 2007.
- [71] Y. Lai and C. Kuo, "A haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Representation*, vol. 11, no. 1, pp. 17–40, 2000.
- [72] B. Lane, "Stereoscopic displays," *Processing and display of three-dimensional data*, pp. 20–32, 1983.
- [73] S. Lee, M. Pattichis, and A. Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 977–992, 2001.
- [74] Y. LeGrand, *Optique physiologique tome troisi me: l'espace visual*. Masson & C, Paris, 1964.

- 
- [75] W. Levelt, *On binocular rivalry*. Mouton, 1968, vol. 2.
- [76] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo, “Kinect identity: Technology and experience,” *Computer*, vol. 44, no. 4, pp. 94–96, 2011.
- [77] W. Li, “Mpeg-4 video verification model version 18.0,” *ISO/IEC JTC1/SC29/WG11, N3908*, 2001.
- [78] A. Liew and H. Yan, “Blocking artifacts suppression in block-coded images using overcomplete wavelet representation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 450–461, 2004.
- [79] W. Lin and C. Jay Kuo, “Perceptual visual quality metrics: A survey,” *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.
- [80] L. Lipton, *Foundations of the stereoscopic cinema: a study in depth*. Van Nostrand Reinhold, 1982, vol. 259.
- [81] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, “Adaptive deblocking filter,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 614–619, 2003.
- [82] Y. Luo and R. Ward, “Removing the blocking artifacts of block-based dct compressed images,” *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 838–842, 2003.
- [83] W. Mark, “Post-rendering 3 d image warping: visibility, reconstruction, and performance for depth-image warping,” Ph.D. dissertation, University of North Carolina at Chapel Hill, 1999.
- [84] D. McAllister, *Stereo computer graphics and other true 3D technologies*. Princeton University Press, 1993.
- [85] M. McFarland, M. Pinson, and S. Wolf, “Batch video quality metric (bvqm) users manual,” *Institute for Telecommunication Sciences, Boulder, Colorado, USA*, 2007 (Last checked: August, 2012). [Online]. Available: <http://www.its.bldrdoc.gov/pub/ntia-rpt/06-441a>.
- [86] L. McMillan Jr, “An image-based approach to three-dimensional computer graphics,” Ph.D. dissertation, Citeseer, 1997.
- [87] A. McNamara, “Visual perception in realistic image synthesis,” in *Computer Graphics Forum*, vol. 20, no. 4. Wiley Online Library, 2001, pp. 211–224.

- 
- [88] D. Meegan, L. Stelmach, and W. Tam, "Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery." *Journal of Experimental Psychology: Applied*, vol. 7, no. 2, p. 143, 2001.
- [89] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, T. Wiegand, *et al.*, "The effects of multiview depth video compression on multiview rendering," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 73–88, 2009.
- [90] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *IEEE International Conference on Image Processing, (ICIP)*, vol. 1. IEEE, September, 2007, pp. 201–204.
- [91] D. Mitchell, "A review of the concept of panum's fusional areas." *American journal of optometry and archives of American Academy of Optometry*, vol. 43, no. 6, p. 387, 1966.
- [92] A. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. Bovik, "Wireless video quality assessment: a study of subjective scores and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 587–599, 2010.
- [93] M. Mrak, T. Zgaljic, and E. Izquierdo, "Influence of downsampling filter characteristics on compression performance in wavelet-based scalable video coding," *Journal of Image Processing, IET*, vol. 2, no. 3, pp. 116–129, 2008.
- [94] M. Mueller, F. Zilly, and P. Kauff, "Adaptive cross-trilateral depth map filtering," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, Tampere, Finland, May, 2010, pp. 1–4.
- [95] J. Neely, "The raf near-point rule," *The British Journal of Ophthalmology*, vol. 40, no. 10, p. 636, 1956.
- [96] A. Neri, P. Campisi, E. Maiorana, and F. Battisti, "3d video enhancement based on human visual system characteristics," pp. 13–15, 2010.
- [97] N. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Transactions on Communications*, vol. 33, no. 6, pp. 551–557, 1985.
- [98] K. N. Ogle and K. Neil, "Researches in binocular vision," vol. 102, 1964.
- [99] B. Oh, H. Wey, and D. Park, "Plane segmentation based intra prediction for depth map coding," in *Picture Coding Symposium (PCS)*. IEEE, Krakow, Poland, May, 2012, pp. 41–44.

- 
- [100] K. Oh, S. Yea, and Y. Ho, “Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video,” in *Picture Coding Symposium, (PCS)*. IEEE, Chicago, Illinois, USA, May, 2009, pp. 233–236.
- [101] K. Oh, S. Yea, A. Vetro, and Y. Ho, “Depth reconstruction filter and down/up sampling for depth coding in 3-d video,” *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 747–750, 2009.
- [102] C. W. Oyster, *The human eye: structure and function*. Sinauer Associates, 1999.
- [103] S. Pastoor, *3D Displays*. John Wiley and Sons, 2005.
- [104] R. Patterson and W. Martin, “Human stereopsis.” *Human Factors*, pp. 669–692, 1992.
- [105] M. Perkins, “Data compression of stereopairs,” *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684–696, 1992.
- [106] M. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies,” in *SPIE Video Communications and Image Processing Conference*, vol. 5150, Santa Clara, CA, USA, January, 2003, pp. 573–582.
- [107] ———, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.
- [108] N. Polak and R. Jones, “Dynamic interactions between accommodation and convergence,” *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 10, pp. 1011–1014, 1990.
- [109] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, “On between-coefficient contrast masking of dct basis functions,” in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics*, vol. 4, January, 2007.
- [110] C. Porac, S. Coren, *et al.*, *Lateral preferences and human behavior*. Springer-Verlag New York, 1981.
- [111] F. Qi, T. Jiang, S. Ma, and D. Zhao, “Quality of experience assessment for stereoscopic images,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, Seoul, Korea, May, 2012, pp. 1712–1715.
- [112] D. Qin, M. Takamatsu, and Y. Nakashima, “Measurement for the panum’s fusional area in retinal fovea using a three-dimension display device,” *Journal of Light & Visual Environment*, vol. 28, no. 3, pp. 126–131, 2004.
- [113] P. Read and M.-P. Meyer, *Restoration of motion picture film*. Butterworth-Heinemann, 2000.

- 
- [114] I. Recommendation, “910, subjective video quality assessment methods for multimedia applications,” *International Telecommunication Union, Geneva, Switzerland*, 1999.
- [115] —, “500-11, methodology for the subjective assessment of the quality of television pictures,” *International Telecommunication Union, Geneva, Switzerland*, 2002.
- [116] —, “144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” *International Telecommunication Union*, 2004.
- [117] —, “264,” *Advanced video coding for generic audiovisual services*, 2009.
- [118] A. Riemens, O. Gangwal, B. Barenbrug, and R. Berretty, “Multi-step joint bilateral depth upsampling,” *Proc. of Electronic Imaging, Visual Communications and Image Processing*, pp. 1–12, January, 2009.
- [119] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, and M. Rupp, “Audiovisual quality estimation for mobile streaming services,” in *2nd International Symposium on Wireless Communication Systems*. IEEE, Siena, Italy, September, 2005, pp. 173–177.
- [120] M. Rosenfield and N. Logan, *Optometry: science, techniques and clinical management*. Butterworth-Heinemann, 2009.
- [121] J. Ruiz-Hidalgo, J. Morros, P. Aflaki, F. Calderero, and F. Marqués, “Multi-view depth coding based on combined color/depth segmentation,” *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 42–52, 2012.
- [122] D. Rusanovskyy and K. Egiazarian, “Video denoising algorithm in sliding 3d dct domain,” in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2005, pp. 618–625.
- [123] S. Rushton and P. Riddell, “Developing visual systems and exposure to virtual reality and stereo displays: some concerns and speculations about the demands on accommodation and vergence,” *Applied Ergonomics*, vol. 30, no. 1, pp. 69–78, 1999.
- [124] G. Saygili, C. Gurler, and A. Tekalp, “Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3d video streaming,” *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 593–601, 2011.
- [125] G. Saygılı, C. Gürler, and A. Tekalp, “Quality assessment of asymmetric stereo video coding,” in *IEEE Int. Conf. on Image Processing (ICIP)*, Hong Kong, September, 2010, pp. 4009–4012.

- 
- [126] Z. Sazzad, S. Yamanaka, Y. Kawayokeita, and Y. Horita, "Stereoscopic image quality prediction," in *International Workshop on Quality of Multimedia Experience, QoMEX*. IEEE, San Diego, California, USA, July, 2009, pp. 180–185.
- [127] C. Schor, "The influence of interactions between accommodation and convergence on the lag of accommodation," *Ophthalmic and Physiological Optics*, vol. 19, no. 2, pp. 134–150, 2002.
- [128] A. Segall and J. Zhao, "Evaluation of texture upsampling with 4-tap cubic-spline filter," *Joint Video Team, Doc. JVT-U042, Hangzhou, China*, 2006.
- [129] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.
- [130] P. Seuntjens, L. Meesters, and W. Ijsselstein, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation," *ACM Transactions on Applied Perception (TAP)*, vol. 3, no. 2, pp. 95–109, 2006.
- [131] J. Shade, S. Gortler, L. He, and R. Szeliski, "Layered depth images," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, Florida, USA, July, 1998, pp. 231–242.
- [132] F. Shao, S. Gu, G. Jang, and M. Yu, "A novel no-reference stereoscopic image quality assessment method," in *IEEE Symposium on Photonics and Optoelectronics (SOPO)*. IEEE, Shanghai, China, May, 2012, pp. 1–4.
- [133] F. Shao, S. Gu, G. Jiang, and M. Yu, "Stereoscopic images quality assessment by jointly evaluating image quality and depth perception," in *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, Sichuan, China, May, 2012, pp. 1963–1966.
- [134] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [135] G. Shen, W. Kim, S. Narang, A. Ortega, J. Lee, and H. Wey, "Edge-adaptive transforms for efficient depth map coding," in *Picture Coding Symposium (PCS)*. IEEE, Nagoya, Japan, December, 2010, pp. 566–569.
- [136] G. Shen, W. Kim, A. Ortega, J. Lee, and H. Wey, "Edge-aware intra prediction for depth-map coding," in *International conference of Image Processing, (ICIP)*, Hong Kong, September, 2010, pp. 3393–3396.



- 
- [137] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding using super-resolution technique," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 6, pp. 755–765, 2011.
- [138] T. Shibata, J. Kim, D. Hoffman, and M. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *Journal of vision*, vol. 11, no. 8, 2011.
- [139] I. Shin and H. Park, "Adaptive up-sampling method using dct for spatial scalability of scalable video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 206–214, 2009.
- [140] K. Shoemake, "Animating rotation with quaternion curves," *ACM SIGGRAPH computer graphics*, vol. 19, no. 3, pp. 245–254, 1985.
- [141] A. Smolic, K. Mueller, P. Merkle, N. Atzpadin, C. Fehn, M. Mueller, O. Schreer, R. Tanger, P. Kauff, and T. Wiegand, "Multi-view video plus depth (mvd) format for advanced 3d video systems," *MPEG and ITU-T SG16 Q*, vol. 6, 2007.
- [142] L. Stelmach, W. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 188–193, 2000.
- [143] L. Stelmach, W. Tam, D. Meegan, A. Vincent, and P. Corriveau, "Human perception of mismatched stereoscopic 3d inputs," in *International Conference on Image Processing, (ICIP)*, vol. 1. IEEE, Vancouver ,Canada, September, 2000, pp. 5–8.
- [144] D. Strohmeier, S. Jumisko-Pyykko, and U. Reiter, "Profiling experienced quality factors of audiovisual 3d perception," in *Second International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, Trondheim, Norway, June, 2010, pp. 70–75.
- [145] D. Strohmeier and G. Tech, "Sharp, bright, three-dimensional: open profiling of quality for mobile 3dtv coding methods," in *Proc. SPIE T*, vol. 75420, March, 2010.
- [146] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Trans. Circuits and Systems for Video Tech*, 2012.
- [147] P. Surman, K. Hopf, I. Sexton, W. Lee, and R. Bates, "Solving the 3d problem, the history and development of viable domestic 3-dimensional video displays," *Three-Dimensional Television: Capture, Transmission, and Display*, 2007.

- 
- [148] R. Suryakumar, “Study of the dynamic interactions between vergence and accommodation,” Ph.D. dissertation, University of Waterloo, 2005.
- [149] T. Suzuki, M. M. Hannuksela, S. Chen, Y. ans Hattori, and G. J. Sullivan, “Mvc extension for inclusion of depth,” *Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1001*, 2012.
- [150] T. Takeda, K. Hashimoto, N. Hiruma, and Y. Fukui, “Characteristics of accommodation toward apparent depth,” *Vision research*, vol. 39, no. 12, pp. 2087–2097, 1999.
- [151] W. J. Tam, G. Alain, L. Zhang, T. Martin, R. Renaud, and D. Wang, “Smoothing depth maps for improved stereoscopic image quality,” in *Proceeding of SPIE*, vol. 5599, December, 2004, p. 163.
- [152] W. Tam, “Image and depth quality of asymmetrically coded stereoscopic video for 3d-tv,” *JVT-W094, San Jose, CA*, 2007.
- [153] Y. Tan and W. Li, “Vision: Trichromatic vision in prosimians,” *Nature*, vol. 402, no. 6757, pp. 36–36, 1999.
- [154] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, “Reference softwares for depth estimation and view synthesis,” *ISO/IEC JTC1/SC29/WG11 M*, vol. 15377, 2008.
- [155] A. Telea, “An image inpainting technique based on the fast marching method,” *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [156] P. Teo and D. Heeger, “Perceptual image distortion,” in *IEEE International Conference on Image Processing*, vol. 2, Austin, Texas, USA, 1994, pp. 982–986.
- [157] D. Tian, D. Graziosi, Y. Wang, N. Cheung, and A. Vetro, “Mitsubishi response to mpeg call for proposal on 3d video coding technology,” *ISO/IEC JTC1/SC29/WG11 MPEG2011/M22663*, 2011.
- [158] F. Toates, “Accommodation function of the human eye,” *Physiological reviews*, vol. 52, no. 4, pp. 828–863, 1972.
- [159] W. Tzschoppe, T. Brueggert, M. Klippstein, I. Relke, U. Hofmann, *et al.*, “Arrangement for two-or three-dimensional display,” 2012, uS Patent 8,238,024.
- [160] I. T. Union, “User requirements for objective perceptual video quality measurements in digital cable television.” ITU Telecommunication Standardization Sector (ITU-T). Geneva, 2000.
- [161] N. Valyrus, “Stereoscopy,” *Focal, London*, 1966.

- 
- [162] C. Van Berkel, “Lenticular screen adaptor,” 2004, uS Patent 6,801,243.
- [163] C. van Berkela and J. A. Clarke, “Characterisation and optimisation of 3d-lcd module design,” in *Proc. of SPIE*, vol. 3012, December, 1997, pp. 179–187.
- [164] C. Vázquez, W. Tam, and F. Speranza, “Stereoscopic imaging: filling disoccluded areas in depth image-based rendering,” in *SPIE*, vol. 6392, March, 2006, pp. 63 920D–12.
- [165] H. Von Helmholtz, *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Voss, 1866.
- [166] G. K. Wallace, “The jpeg still picture compression standard,” *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.
- [167] B. Wandell, *Foundations of vision*. Sinauer Associates, 1995.
- [168] X. Wang, M. Yu, Y. Yang, and G. Jiang, “Research on subjective stereoscopic image quality assessment,” in *Proc. SPIE*, vol. 7255, no. 725509, 2009, pp. 725 509–1–725 509–10.
- [169] Z. Wang and A. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.
- [170] Z. Wang, A. Bovik, and L. Lu, “Why is image quality assessment so difficult?” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, Orlando, Florida, USA, May, 2002, pp. IV–3313.
- [171] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [172] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2. IEEE, 2003, pp. 1398–1402.
- [173] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2012.
- [174] C. Wheatstone, “Contributions to the physiology of vision.—part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision,” *Philosophical transactions of the Royal Society of London*, vol. 128, pp. 371–394, 1838.
- [175] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

- 
- [176] M. Wildeboer, T. Yendo, M. Tehrani, T. Fujii, and M. Tanimoto, “Color based depth up-sampling for depth compression,” in *Picture Coding Symposium (PCS)*. IEEE, Nagoya, Japan, December, 2010, pp. 170–173.
- [177] H. Wilson, R. Blake, and S. Lee, “Dynamics of travelling waves in visual perception,” *Nature*, vol. 412, no. 6850, pp. 907–910, 2001.
- [178] S. Winkler, “Perceptual distortion metric for digital color video,” in *Electronic Imaging*. International Society for Optics and Photonics, 1999, pp. 175–184.
- [179] —, *Digital video quality*. Wiley, 2005.
- [180] —, “Perceptual video quality metrics—a review,” pp. 155–172, 2005.
- [181] S. Winkler and P. Mohandas, “The evolution of video quality measurement: from psnr to hybrid metrics,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.
- [182] H. Wu and K. Rao, *Digital video image quality and perceptual coding*. CRC, 2005, vol. 28.
- [183] C. Xiao, Y. Nie, W. Hua, and W. Zheng, “Fast multi-scale joint bilateral texture upsampling,” *The Visual Computer*, vol. 26, no. 4, pp. 263–275, 2010.
- [184] Z. Xiong, M. Orchard, and Y. Zhang, “A deblocking algorithm for jpeg compressed images using overcomplete wavelet representations,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 2, pp. 433–437, 1997.
- [185] Y. Yang, N. Galatsanos, and A. Katsaggelos, “Regularized reconstruction to reduce blocking artifacts of block discrete cosine transform compressed images,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 6, pp. 421–432, 1993.
- [186] S. Yasakethu, W. Fernando, B. Kamolrat, and A. Kondoz, “Analyzing perceptual attributes of 3d video,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 864–872, 2009.
- [187] J. You, L. Xing, A. Perkis, and X. Wang, “Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis,” in *Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, USA, January, 2010.
- [188] M. Yuen and H. Wu, “A survey of hybrid mc/dpcm/dct video coding distortions,” *Signal processing*, vol. 70, no. 3, pp. 247–278, 1998.

- 
- [189] A. Zakhor, “Iterative procedures for reduction of blocking effects in transform image coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 2, no. 1, pp. 91–95, 1992.
- [190] J. Zhu, L. Wang, R. Yang, and J. Davis, “Fusion of time-of-flight depth and stereo for high accuracy depth maps,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Anchorage, Alaska, USA, June, 2008, pp. 1–8.
- [191] S. Zinger, L. Do, *et al.*, “Free-viewpoint depth image based rendering,” *Journal of Visual Communication and Image Representation*, vol. 21, no. 5, pp. 533–541, 2010.

# Appendix - Publications



- [P1] **P. Aflaki**, M. M. Hannuksela, D. Rusanovskyy, and M. Gabbouj, “Non-linear depth map resampling for depth-enhanced 3D video coding,” IEEE Signal Processing Letters, Vol. 20, issue 1, pp. 87-90, January, 2013.

© IEEE, 2013, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.





# Non-Linear Depth Map Resampling for Depth-Enhanced 3D Video Coding

Payman Aflaki<sup>a</sup>, Miska M. Hannuksela<sup>b</sup>, Dmytro Rusanovskyy<sup>b</sup>, Moncef Gabbouj<sup>a</sup>

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

**Abstract**— Depth-enhanced 3D video coding includes coding of texture views and associated depth maps. It has been observed that coding of depth map at reduced resolution provides better rate-distortion performance on synthesized views comparing to utilization of full resolution (FR) depth maps in many coding scenarios based on the Advanced Video Coding (H.264/AVC) standard. Conventional techniques for down and upsampling do not take typical characteristics of depth maps, such as distinct edges and smooth regions within depth objects, into account. Hence, more efficient down and upsampling tools, capable of preserving edges better, are needed. In this letter, novel non-linear methods to down and upsample depth maps are presented. Bitrate comparison of synthesized views, including texture and depth map bitstreams, is presented against a conventional linear resampling algorithm. Objective results show an average bitrate reduction of 5.29% and 3.31% for the proposed down and upsampling methods with ratio  $\frac{1}{2}$ , respectively, comparing to the anchor method. Moreover, a joint utilization of the proposed down and upsampling brings up to 20% and on average 7.35% bitrate reduction.

**Index Terms**—MVC, depth map, resampling, non-linear.

## I. INTRODUCTION

The multiview video plus depth (MVD) format [1], where each video data pixel is associated with a corresponding depth map value, is one of the most promising methods for providing 3D video services flexible for different types of multiview displays as well as user adaptation at disparity between rendered views. The MVD format allows reducing the input data for the 3DV systems significantly, since most of the views will be rendered from the available decoded views and depth maps using a Depth Image Based Rendering (DIBR) [2] algorithm.

3D video coding (3DV) standardization by the Moving Picture Experts Group (MPEG) is a recent activity targeting at enabling a variety of display types and preferences including varying baseline to adjust the depth perception. Another important target of the MPEG 3DV standardization is the support for multiview auto-stereoscopic displays, thus many high-quality views shall be available in decoder/display side prior to displaying. As the existing video compression standards were found to be sub-optimal to achieve these targets, MPEG issued a Call for Proposals for 3D video coding (hereafter referred to as the 3DV CfP) [3] to kick off the 3DV standardization activity targeting to provide 3D

enhancement to the existing the Multiview Video Coding extension of the Advanced Video Coding standard, H.264/MVC [4], as well as to the ongoing High Efficiency Video Coding (HEVC) standardization. As one consequence of the 3DV CfP, a H.264/MVC-based test model [5] (hereafter referred to as 3DV-ATM) was chosen and has been further developed by MPEG as collaborative standardization effort. In addition to exploiting temporal and inter-view correlation among texture or depth views to achieve high coding efficiency, 3DV-ATM provides means to encode depth maps into the same bitstream with texture and enhances H.264/MVC with coding tools utilizing the correlation between depth and texture data.

In 3D video applications, depth maps are used for synthesizing new images but not to be directly viewed by end users. Thus, when coding depth maps, the goal is to maximize the perceived visual quality of the rendered virtual color views instead of the visual quality of the depth maps themselves [6]. Traditional video coding methods have been designed to operate through a Rate-Distortion Optimization (RDO) of coded data and a pixel-based distortion introduced by codec, e.g. Sum of Absolute Differences (SAD). However, coding distortions of a depth map typically have a non-linear impact on the visual quality of rendered views [7]. For example, errors in the depth map close to a sharp edge can result in severe rendering artifacts, while errors on a smooth area may have negligible subjective influence on the final quality. Therefore, utilization of traditional RDO for depth map compression may result in suboptimal performance of a 3D video coding system [7].

As demonstrated in many of the responses to the 3DV CfP and enabled in 3DV-ATM, coding of depth map data at a reduced resolution is a viable solution for improving the rate-distortion performance of the complete 3D video coding system. In such systems, depth map data is downsampled prior to the encoding and upsampled to the original FR after decoding. Obviously, downsampling of depth maps, which is performed in combination with low-pass filtering for aliasing suppression, may lead to smoothed edges and therefore to a significant distortion in rendered views.

In this letter, a novel algorithm for non-linear down and upsampling of depth map is presented. The proposed

algorithm preserves edges in processed depth map data and provides quality improvement in synthesized images.

The rest of the letter is organized as follows. Section II provides a review of depth map resampling methods. The proposed down and upsampling methods are introduced in section III while the simulation setup and results are presented in section IV. Finally, the letter concludes in section V.

## II. DEPTH MAP RESAMPLING

Downsampling traditionally includes low pass filtering, which suppresses high frequency components in the depth map and therefore leads to over-smooth edges. The consequent quality reduction due to resampling causes significant visual artifacts in synthesized views particularly at object boundaries. Hence, edge-preserving downsampling for depth map should be considered even though traditional image downsampling techniques use linear filters not designed to preserve edges. For example, in [8] and [9], the median value of an  $N \times N$  window was chosen as the most representative value to be used at reduced-resolution depth map (where factor  $N$  specifies the downsampling ratio).

Similarly to downsampling, upsampling should preserve depth edges. In various works, e.g. [9] and [10], cross-component bi-lateral filtering has been used for depth upsampling. In a cross-component bi-lateral filter, the similarity of co-located texture samples is used to derive filter weights for depth in addition to the conventional filtering window applied spatially for the depth samples.

Another approach for coded depth upsampling and restoration was used in [8] and [11]. Other than a depth resampling technique to improve the quality of rendered views, authors proposed that a decoded low resolution depth map image to be processed with a depth reconstruction filter. This filter consists of a novel frequent-low-high filter and a bilateral filter. Depth map is first upsampled using a nearest neighbor filter, which is followed by post-processing using a median filter, a frequent-low-high filter and a cross-component bi-lateral filter. The 2D median filter is used to smooth blocking artifacts caused by coding. The frequent-low-high filter is a non-linear filter used to recover object boundaries, which results into selecting either the most frequently occurring sample value below or above the median sample value within a filter window. The bilateral filter is used to eliminate the errors still present after both filtering procedures.

In [12] an edge adaptive upsampling method for better compression of depth maps is presented. In this work edge information is extracted from the high resolution reconstructed texture video by applying  $3 \times 3$  Sobel filter operators. Gradients caused by texture transitions, rather than depth changes, are eliminated by considering the local depth intensity gradients. Then the linear interpolation filters are replaced with a locally adaptive filter. Test results reported in [12], show that the proposed technique outperformed linear MPEG upsampling filter [13] in terms of objective and subjective quality of synthesized views. However, the utilization of texture data in upsampling process of depth map can be considered a drawback of the proposed method due to a significant increase in the memory access bandwidth and computational complexity.

## III. PROPOSED DOWN AND UP SAMPLING METHODS

The proposed down and upsampling method presented in this section can be applied directly to depth maps and do not need complementary information from the reconstructed or the decoded texture images. In following sub-sections a detailed description of the algorithms is presented.

### A. Downsampling

To perform the proposed downsampling method, a block of pixels (BOP) will be determined based on the downsampling ratio. The FR image will be covered with the necessary number of non-overlapping BOPs and for each BOP a single value will be calculated to present it in the downsampled image. The size of the BOP is defined as the reciprocal of the downsampling ratio; e.g. if the image is downsampled with ratios  $1/x$  and  $1/y$  (both  $x$  and  $y$  are positive values equal or bigger than 1) along the horizontal and vertical direction where the size of the BOP is specified with  $x$  and  $y$  in width and height, respectively.

The proposed downsampling method utilizes a closeness-favored averaging algorithm as described in the following paragraphs. In the first step an average over the BOP will be calculated, as seen in (1).

$$Avg_{BOP} = \frac{\sum_{i=1}^x \sum_{j=1}^y BOP_{(i,j)}}{x \times y} \quad (1)$$

where  $BOP_{(i,j)}$  presents a pixel value where  $i$  and  $j$  are the horizontal and vertical pixel indices within the BOP.

In the next step pixels of BOP are categorized into two sets as shown in (2).

$$BOP_{(i,j)} \in \begin{cases} G_{high}, & \text{if } BOP_{(i,j)} \geq Avg_{BOP} \\ G_{low}, & \text{otherwise.} \end{cases} \quad (2)$$

where  $BOP_{(i,j)}$  is the same as in equation (1).

If the number of pixels in  $G_{high}$  is equal to or greater than half of the number of pixels in the BOP, the Estimated Value ( $EV$ ) of the associated BOP is an average over the pixel values of  $G_{high}$ . Otherwise,  $EV$  is set to  $Avg_{BOP}$ , as shown in (3) and (4).

$$Avg_{G_{high}} = \frac{\sum_{i=1}^x \sum_{j=1}^y BOP_{(i,j)}}{Count(G_{high})}, \quad BOP_{(i,j)} \in G_{high} \quad (3)$$

$$EV = \begin{cases} Avg_{G_{high}}, & Count(G_{high}) \geq \frac{Count(BOP)}{2} \\ Avg_{BOP}, & \text{otherwise} \end{cases} \quad (4)$$

where  $Count(X)$  counts the number of elements in  $X$ . The calculated  $EV$  is the value which represents the considered BOP in the downsampled image. As can be observed from the equations, if at least half of the pixels in a BOP are classified to belong to objects that are close-by, i.e. closer than the average depth value of the BOP, the method considers only those pixels in downsampling and hence attempts to preserve sharp boundaries of foreground objects. Since the entire FR image is processed with non-overlapped

BOPs, the calculated EVs form a downsampled version of the input image.

### B. Upsampling

Considering Figure 1 pixel values  $\{A, B, C, D, E, F, G, H, I\}$  in the downsampled image are utilized to upsample pixel  $E$  and calculate values of pixels  $\{a, b, c, d\}$  in the associated BOP in the upsampled image. Afterwards,  $a, b, c,$  and  $d$  will be utilized to create possible remaining pixel values in the BOP of upsampled image.

Let us consider the pixel which needs to be upsampled (pixel  $E$  in Figure 1). To calculate the value of the top-left pixel in the BOP of the upsampled image ( $a$  in Figure 1), the pixel values on the left and top of  $E$  will be considered ( $D$  and  $B$  in Figure 1, respectively).

In the first step, the absolute differences of  $E$  with  $D$  and  $B$  are calculated. This is shown in (5) and (6).

$$diff_{EB} = |E - B| \quad (5)$$

$$diff_{ED} = |E - D| \quad (6)$$

The filter window (FW) by which the value of  $a$  will be calculated is defined as following. If both  $diff_{EB}$  and  $diff_{ED}$  are smaller than a threshold ( $th$ ), then it is assumed that  $A, B, D,$  and  $E$  belong to the same depth region, and consequently the final FW contains pixels  $A, B, D,$  and  $E$ . Otherwise, the final FW is chosen to contain only  $A, B,$  and  $D$ , as shown in (7). This choice of the filter window attempts to restore the shape of a depth boundary between  $E$  and a depth object containing  $A, B,$  and  $D$ .

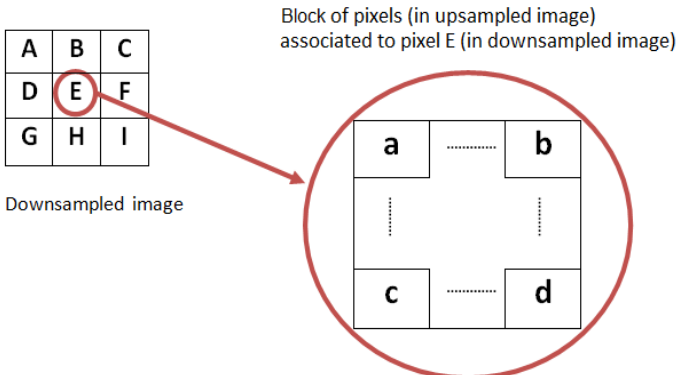
$$FW \in \begin{cases} \{A, B, D, E\}, & diff_{EB} < th \text{ and } diff_{ED} < th \\ \{A, B, D\}, & otherwise \end{cases} \quad (7)$$

In the next step, the average of pixel values in selected  $FW$  is calculated and utilized as  $a$  in the upsampled image (see Figure 1). This is presented in (8).

$$a = average(FW) \quad (8)$$

The complete procedure to calculate  $a$  from  $A, B, D,$  and  $E$  can be presented with function *upsampler* as shown in (9).

$$a = upsampler(E, A, B, D) \quad (9)$$



**Figure 1.** To be upsampled pixel value ( $E$ ) and associated block of pixels in upsampled image

The same process is applied for the other three corner pixels, i.e.  $b, c,$  and  $d$ . The pixel values of  $b, c,$  and  $d$  in Figure 1 can be calculated using the equations shown in (10).

$$\begin{aligned} b &= upsampler(E, C, B, F) \\ c &= upsampler(E, G, D, H) \\ d &= upsampler(E, I, H, F) \end{aligned} \quad (10)$$

Finally, when  $a, b, c,$  and  $d$  in the BOP of the upsampled image are available, bi-linear interpolation is applied to obtain the remaining pixel values in the BOP.

## IV. SIMULATIONS

### A. Simulation Setup

3DV-ATM reference software [5] (hereafter referred to as reference software (RS)) was utilized for encoding the multiview plus depth (MVD) data. Simulations were conducted according to the MPEG 3DV Common Test Conditions (CTC) [14].

Depth maps for all resampling schemes were first downsampled to half resolution along each of coordinate axes prior to encoding and upsampled to the FR after decoding. The threshold utilized for proposed upsampling process was fixed to 16 for all sequences. The view synthesis was performed with VSRS software, version 3.5 [15] with configuration and camera parameters information provided with MPEG 3DV CTC [14]. In our experiment, we provide the results for C3 scenario, described in [14] where three evenly distributed intermediate views between each two input (coded) views were synthesized.

### B. Simulation Results

The proposed algorithm was tested against the depth map resampling utilized in the RS, with 12-tap low pass filtering in downsampling according to Joint Scalable Video Model (JSVM) [16] and bi-linear upsampling for upsampling. The performance of the proposed down and upsampling methods was evaluated separately against the techniques utilized in the RS with the two following set of experiments:

- First experiment: A combination of the proposed downsampling and RS upsampling compared against RS used for both downsampling and upsampling
- Second experiment: A combination of the RS downsampling and the proposed upsampling compared against RS used for both downsampling and upsampling

In the third experiment the efficiency of the method in [11] and a joint utilization of the proposed down and upsampling was tested against the RS.

Simulation results for the first and second experiments using Bjontegaard delta bitrate and delta Peak Signal-to-Noise Ratio (PSNR) [17] are reported in Tables I while results of the third experiment are presented in table II. In these calculations the total bitrate of texture plus depth maps along with the average luma PSNR of all six synthesized views were considered.

Table I shows that the proposed downsampling method outperformed the anchor method of [14] by 5.29% of Bjontegaard delta bitrate reduction (dBR) and 3.31% dBR on average was achieved by the proposed upsampling algorithm. Results of the third experiment show that a joint utilization of both proposed methods provided 7.35% dBR comparing the RS. From our simulations, the algorithm presented in [11] performed worse than anchor objectively. However, based on our expert subjective viewing, the perceived quality of our proposed method and the algorithm presented in [11] outperformed that of RS. Moreover, in [11] it is claimed that by applying the proposed filter on depth maps a better compression for depth map and higher subjective quality for rendered views are achieved. Additionally, the decoder execution time of the proposed method was 85% of the RS on average, while the method presented in [11] has more computation operations per pixel than our proposed algorithm.

## V. CONCLUSIONS

Due to the characteristics of depth maps, coding of depth maps at a lower spatial resolution than the resolution of luma texture pictures typically results into improved rate-distortion performance. However, traditional resampling algorithms which use linear filtering result to significant distortions introduced to rendered views. In this experiment, we improved the depth-enhanced 3D video coding through edge-

TABLE I. FIRST AND SECOND EXPERIMENTS: PERFORMANCE OF PROPOSED DOWN AND UP SAMPLING AGAINST ANCHOR

	Proposed downsampling against anchor		Proposed upsampling against anchor	
	dBR,%	dPSNR ,dB	dBR,%	dPSNR ,dB
Poznan Hall2	-2.47	0.08	-1.61	0.05
Poznan Street	-3.43	0.10	-1.93	0.05
Undo Dancer	-17.15	0.65	-9.87	0.33
Ghost Town	-5.69	0.21	-1.88	0.07
Kendo	-2.62	0.12	-2.62	0.12
Balloons	-1.07	0.05	-1.01	0.04
Newspaper	-4.57	0.17	-4.21	0.16
<b>Average</b>	<b>-5.29</b>	<b>0.20</b>	<b>-3.31</b>	<b>0.12</b>

TABLE II. THIRD EXPERIMENT: PERFORMANCE OF JOINT UTILIZATION OF PROPOSED DOWN/UP SAMPLING AGAINST RS

	Method presented in [11]		Proposed method	
	dBR,%	dPSNR ,dB	dBR,%	dPSNR ,dB
Poznan Hall2	0.22	-0.27	-4.34	0.15
Poznan Street	0.79	-0.02	-5.24	0.15
Undo Dancer	0.99	-0.04	-20.40	0.87
Ghost Town	1.91	-0.07	-8.54	0.33
Kendo	1.72	-0.07	-4.53	0.21
Balloons	1.92	-0.09	-1.19	0.05
Newspaper	2.94	-0.1	-7.19	0.29
<b>Average</b>	<b>1.50</b>	<b>-0.09</b>	<b>-7.35</b>	<b>0.29</b>

preserving techniques for depth map resampling. Two novel algorithms for down and upsampling depth maps were presented in this letter. Results showed that proposed down and upsampling steps with ratio  $\frac{1}{2}$  outperform MPEG 3DV anchor resampling methods by 7.35% of dBR on average and up to 20.4%. In addition to this, the proposed implementation decreased the decoder execution time by 15% compared to the MPEG H.264/AVC-based 3DV reference software.

## ACKNOWLEDGMENT

The authors would like to thank Prof. M. Domański, et al. for providing Poznan sequences and Camera Parameters [18].

## REFERENCES

- [1] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," Proc. of International Conf. on Image Processing, vol. 1, pp. 201-204, Oct. 2007.
- [2] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," Proc. of SPIE stereoscopic displays and virtual reality systems XI, pp. 93-104, Jan. 2004.
- [3] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, March 2011.
- [4] "Advanced video coding for generic audiovisual services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2012.
- [5] "Test model for AVC based 3D video coding," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12558, Feb. 2012.
- [6] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, T. Wiegand, "The effects of multiview depth video compression on multiview rendering," Signal Processing: Image Communication, vol. 24, pp. 73-88, Jan. 2009.
- [7] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," Proc. of IS&T/SPIE Electronic Imaging, vol. 7543, pp. 75430B-75430B-10, Jan. 2010.
- [8] D. Tian, D. Graziosi, Y. Wang, N. Cheung, A. Vetro, "Mitsubishi Response to MPEG Call for Proposal on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/M22663, Nov. 2011.
- [9] M. O. Wildeboer, T. Yendo, M. Panahpour Tehrani, T. Fujii, M. Tanimoto, "Color Based Depth Upsampling for Depth Compression," Proc. of Picture Coding Symposium, pp. 170-173, Dec. 2010.
- [10] A. K. Riemens, O. P. Gangwal, B. Barenbrug, R-P. M. Berretty, "Multi-step joint bilateral depth upsampling," Proc. of Visual Communications and Image Processing, vol. 7257, pp. 72570M-72570M-12 Jan. 2009.
- [11] K.-J. Oh, S. Yea, A. Vetro, and Y.-S. Ho, B, "Depth reconstruction filter and down/up sampling for depth coding in 3-D video," IEEE Signal Process. Letters, vol. 16, no. 9, pp. 747-750, Sep. 2009.
- [12] E. Ekmekcioglu, M. Mrak, S. Worrall, and A. M. Kondoz, "Utilisation of edge adaptive upsampling in compression of depth map videos for enhanced free-viewpoint rendering," Proc. of International Conf. on Image Processing, pp. 733-736, Nov. 2009.
- [13] G. Sullivan and S. Sun, "Spatial Scalability Filters," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6/JVT-P007, July 2005
- [14] "Common test conditions for 3DV experimentation," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12560, Feb. 2012.
- [15] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," ISO/IEC JTC1/SC29/WG11/M15377, Apr. 2008.
- [16] JM reference software: <http://iphome.hhi.de/suehring/tml/download>
- [17] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," ITU-T SG16 Q.6 document VCEG-M33, April 2001.
- [18] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan multiview video test sequences and camera parameters," ISO/IEC JTC1/SC29/WG11 MPEG2009/M17050, Oct. 2009.

[P2] **P. Aflaki**, M. M. Hannuksela, H. Sarbolandi, and M. Gabbouj, “Simultaneous 2D and 3D perception for stereoscopic displays based on polarized or active shutter glasses,” Elsevier Journal of Visual Communication and Image Representation, 2013.

© Elsevier, 2013, Reprinted with permission.

# Simultaneous 2D and 3D perception for stereoscopic displays based on polarized or active shutter glasses

Payman Aflaki<sup>a</sup>, Miska M. Hannuksela<sup>b</sup>, Hamed Sarbolandi<sup>a</sup>, Moncef Gabbouj<sup>a</sup>

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland

<sup>b</sup>Nokia Research Center, Tampere, Finland

**Abstract**— Viewing stereoscopic 3D content is typically enabled either by using polarizing or active shutter glasses. In certain cases, some viewers may not wear viewing glasses and hence, it would be desirable to tune the stereoscopic 3D content so that it could be simultaneously watched with and without viewing glasses. In this paper we propose a video post-processing technique which enables good quality 3D and 2D perception of the same content. This is done through manipulation of one view by making it more similar to the other view to reduce the ghosting artifact perceived without viewing glasses while 3D perception is maintained. The proposed technique includes three steps: disparity selection, contrast adjustment, and low-pass filtering. The proposed approach was evaluated through an extensive series of subjective tests, which also revealed good adjustment parameters to suit viewing with and without viewing glasses with an acceptable 3D and 2D quality, respectively.

**Index Terms**—Stereoscopic; depth perception; subjective quality assessment; 3DV; 2DV.

## 1. INTRODUCTION

In the recent years, the number of 3D movie titles has increased considerably both at cinemas and as Blu-ray 3D discs. Moreover, broadcast of stereoscopic video content is provided commercially on a few television channels. Hence, many user side devices are already capable of processing stereoscopic 3D content whose volume is expected to rise sharply in the coming years. Preferences of customers drive the direction of improvements and novelties in different presentation methods of the 3D content and it is therefore important to understand the habits of viewing 3D content and mechanisms of the human vision. Psycho-visual aspects must therefore be considered when displaying 3D content.

The human vision system (HVS) perceives color images using receptors on the retina of the eye which respond to three broad color bands in the regions of red, green and blue in the color spectrum. HVS is more sensitive to overall luminance changes than to color changes. The major challenge in understanding and modeling visual perception is that what people see is not simply a translation of retinal stimuli (i.e., the image on the retina). Moreover, HVS has a limited sensitivity; it does not react to small stimuli, it is not able to discriminate between signals with an

infinite precision, and it also presents saturation effects. In general one could say it achieves a compression process in order to keep visual stimuli for the brain in an interpretable range.

Stereoscopic vision is one of the principal methods by which humans extract 3D information from a scene. HVS is able to fuse the sensory information from the two eyes in such a way that a 3D perception of the scene is formed in a process called stereopsis. In stereoscopic presentation, the brain registers slight perspective differences between left and right views to create a 3D representation incorporating both views. In other words, the visual cortex receives information from each eye and combines this information to form a single stereoscopic image. Presenting different views for each eye (stereoscopic presentation) usually results into binocular rivalry where the two monocular patterns are perceived alternately [1]. In such a case, where dissimilar monocular stimuli are presented to corresponding retinal locations of the two eyes, rather than perceiving stable single stimuli, two stimuli compete for perceptual dominance. Rivalry can be triggered by very simple stimulus differences or by differences between complex images. These include differences in color, luminance, contrast polarity, form, size, and velocity. Stronger, high-contrast stimuli lead to stronger perceptual competition. In particular cases, one of the two stimuli dominates the field. This effect is known as binocular suppression [2], [3]. It is assumed according to the binocular suppression theory that the HVS fuses the two images with different levels of sharpness such that the perceived quality is close to that of the sharper view [4]. In contrast, if both views show different amounts of blocking artifacts, no considerable binocular suppression is observed and the binocular quality of a stereoscopic sequence is rated close to the mean quality of both views [5].

Binocular suppression has been exploited in asymmetric stereoscopic video coding, for example by providing one of the views with lower spatial resolution [6] or with lower frequency bandwidth [7], fewer color quantization steps [8], or coarser transform-domain quantization [9], [10]. In this paper we exploit binocular suppression and asymmetric quality between views in

another domain, namely presentation of stereoscopic 3D content simultaneously on a single display for viewers with and without viewing glasses. Such a viewing situation may occur, for example, when television viewing is not active, but the television set is just being kept on as a habit. The television may be located in a central place at home, where many family members are spending their free time. Consequently, there might be viewers actively watching the television with glasses and while others are primarily doing something else (without glasses) and just momentarily peeking at the television. Furthermore, the price of the glasses, particularly the active ones, might constrain the number of glasses households are willing to buy. Hence, in some occasions, households might not have a sufficient number of glasses for family members and visitors watching the television. While glasses-based stereoscopic display systems provide a good stereoscopic viewing quality, the perceived quality of the stereo picture or picture sequence viewed without glasses is intolerable. Recently, authors in [11] presented a system for automatic 2D/3D display mode selection based on whether the users in front of the 3D display wear viewing glasses. In the research presented in [11] a combination of special viewing glasses and a camera on top of the display enables such display mode selection. However, this approach does not solve the problem of a mixed group of observers, some with and some without viewing glasses and only enables switching between 2D and 3D presentation based on the number of subjects with or without viewing glasses in front of the display.

We enable the same content to be simultaneously viewed both in 3D with viewing glasses and in 2D without viewing glasses by digital signal processing of the decoded stereoscopic video content, making the perceived quality in glasses-based stereoscopic viewing systems acceptable for viewers with and without 3D viewing glasses simultaneously. Viewers with glasses should be able to perceive stereoscopic pictures with acceptable quality and good depth perception, while viewers without glasses should be able to perceive single-view pictures i.e. one of the views of the stereoscopic video. The proposed processing is intended to take place at the display and can be adapted for example based on the ratio of users with and without viewing glasses. In the proposed algorithm, one of the views is processed so that its presence becomes harder to perceive when viewing the content without viewing glasses, while the quality and 3D perception is not compromised much thanks to binocular suppression. The proposed method includes three steps, namely disparity adaptation, low-pass filtering of the non-dominant view, and contrast adjustment. While known methods are used for each processing step, we are not aware of previous research works tackling the same problem, i.e. stereoscopic 3D content being simultaneously viewed with viewing glasses by some users and without viewing glasses by other users.

The rest of this paper is organized as follows. In section 2 we present a literature review of the research fields related to the algorithm proposed in the paper, while the

proposed post-processing algorithm is described in section 3. Test setup and results are presented in sections 4 and 5, respectively. Finally the paper concludes in section 6.

## 2. LITERATURE REVIEW

In this section, we provide an extensive literature review focused on the operation of human visual system when observing an asymmetric quality stereoscopic video. Different types of asymmetry are classified and subjective assessment results are reported in sections 2.1 and 2.2 from perception and video compression viewpoints, respectively. Moreover, in section 2.3, we discuss the effect of camera separation on the depth perception. These techniques provide a basis for rendering algorithms utilized in this study. In section 2.4 we summarize some key aspects affecting the perceived 3D video quality, which are subsequently taken into consideration in the performed subjective viewing experiment. Finally, in section 2.5, the concept of depth-enhanced multiview video coding is described, as it can provide an unlimited number of rendered views at the 3D display. This coding approach can be exploited to display stereoscopic video with arbitrary camera separations, hence facilitating the disparity adaptation step of the method proposed in this paper.

### 2.1. Visual perception of asymmetric stereoscopic video

Binocular suppression provides an opportunity to use different types of asymmetry between views. Many research works have been carried out to study which types of asymmetry are subjectively most pleasing to human observers or closest to the symmetric stereoscopic video and to find optimal settings for various parameters related to the strength of asymmetry.

Typically the greater the amount of high frequency components (more detail), the better the 3D perception of the objects. This means that the stereo acuity decreases when the amount of blurring increases [12]. However, [13] studied this topic in more detail showing that within certain limits, it is possible to perceive stimuli well in 3D even when one eye sees a blurred image while the other eye sees a sharper one.

The capability of the HVS to fuse stereo pairs of different sharpness has been studied in many papers. Authors in [6] subjectively assessed the quality of uncompressed mixed-resolution asymmetric stereoscopic video by downsampling one view with ratios 1/2, 3/8, and 1/4. The results show that while downsampling ratio is equal to 1/2 the average subjective score has sufficient subjective quality which is comparable to that of full resolution stereo pair. A similar experiment was conducted by Stelmach in [14] where the response of HVS to mixed-resolution stereo video sequences where one view was low-pass filtered was explored by performing a series of subjective tests. Subjects rated the overall quality, sharpness, and depth perception of stereo video clips. The results show that the overall sensation of depth was unaffected by low-pass filtering, while ratings of quality and sharpness were strongly



weighted towards the eye with the greater spatial resolution. Moreover, authors in [7] evaluated the perceptual impact of low-pass filtering applied to one view of a stereo image pairs and stereoscopic video sequences in order to achieve an asymmetric stereo scenario. The results showed that binocular perception was dominated by the high quality view when the other view was low-pass filtered.

## 2.2. Asymmetric stereoscopic video coding

The types of asymmetric video coding can be coarsely classified into mixed-resolution, asymmetric sample-domain quantization, asymmetric transform-domain quantization and asymmetric temporal resolution. Furthermore, a combination of different types of scalabilities can be used. The different types of asymmetric stereoscopic video coding are reviewed briefly in the sequel.

Mixed-resolution stereoscopic video coding [15], also referred to as resolution-asymmetric stereoscopic video coding, introduces asymmetry between views by low-pass filtering one view and hence providing smaller amount of spatial details or a lower spatial resolution. Furthermore, usually a coarser sampling grid is utilized for the low-pass-filtered image, i.e. the content is represented with fewer pixels. Mixed-resolution coding can also be applied for a subset of color components. For example, in [16], luma pictures of both views had equal resolution while chroma pictures of one view were represented by fewer samples than the respective chroma pictures of the other view.

In asymmetric transform-domain quantization the transform coefficients of the two views are quantized with a different step size. As a result, one of the views has a lower fidelity and may be subject to a greater amount of visible coding artifacts, such as blocking and ringing. In [9], the authors performed a series of subjective test experiments on coded stereoscopic video clips with asymmetric luminance qualities. Asymmetric luminance was achieved with coarser quantization of transform coefficient values in one luma view. Subjective results show that stereoscopic video coding with asymmetric luminance information achieved a bitrate reduction from 9% to 34% while maintaining the just noticeable distortion as introduced in [17]. Moreover, authors in [10] subjectively compared the quality of coded mixed-resolution stereoscopic video with that of compressed full-resolution video. The results revealed that under the same bitrate constraint, the same subjective quality can be expected while decreasing the spatial resolution of one view by a factor of 1/2 horizontally and vertically.

In asymmetric sample-domain quantization [8] the sample values of each view are quantized with a different step size. A higher compression ratio can be achieved for the quantized view compared to the other view, due to fewer quantization steps. Both luma and chroma samples can be processed with different quantization step sizes. If the number of quantization steps in each view matches a power of two, a special case of asymmetric sample-domain quantization, called bit-depth-asymmetric stereoscopic video, can be achieved. [8] presents a video coding scheme

based on uneven quantization steps for luma sample values of left and right views along with spatial downsampling. Results of subjective quality assessment showed that the average ratings of proposed method outperformed full resolution symmetric and mixed resolution asymmetric stereoscopic video coding schemes with different downsampling ratios.

To our knowledge, asymmetric contrast has not been utilized in stereoscopic video compression. However, authors in [18] subjectively assessed the subjective quality of a wide range of binocular image imperfections by pointing out asymmetry threshold values which provide equal visual comfort. It was found that the contrast difference between views should not exceed 25% to prevent eye strain in subjects.

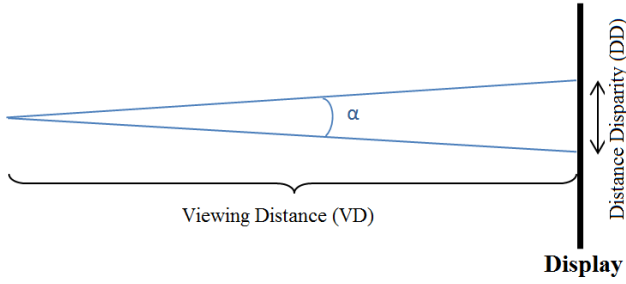
## 2.3. Impact of parallax on depth perception

Screen parallax is created by the difference between the left and right eye images on the 3D display. We need to converge and accommodate (focus) the eyes in order to project the object of interest to the fovea in both eyes. The distance between us and the object of interest defines the amount of convergence and accommodation in our eyes. Convergence can be defined as a process that is basically disparity driven and consists of the movement of the two eyes in opposite direction to locate correctly the area of interest on the fovea. Accommodation tries to remove blur and hence, alters the lens to focus the area of interest on the fovea [19].

Under natural conditions the accommodation and convergence systems are reflexively linked. The amount of accommodation needed to focus on an object changes proportionally to the amount of convergence required to project the same object on the fovea of the eyes. Under conditions of binocular fusion, for a certain amount of convergence, accommodation has a certain depth of focus, in which it can move freely and objects are perceived properly [20].

An area defining an absolute limit for disparities that can be fused in HVS is known as Panum's fusional area [21], [22]. It describes an area, within which different points projected on to the left and right retinas produce binocular fusion and sensation of depth. Hence, horizontal disparity should be limited within Panum's fusional area. Otherwise, excessive disparity could cause double vision or severe visual fatigue. The limits of Panum's fusional area are affected by many factors e.g. including stimulus size, spatial frequency, exposure duration, temporal effects, continuous features, and amount of luminance [21]. Disparities beyond 60 to 70 arcmin are assumed to cause visual discomfort and eye strain [23], [24].

Camera separation creates a disparity between the same object on the left- and right-view images on a display, which can be expressed in terms of number of pixels. Based on the display width and resolution, the disparity can be converted from a number of pixels to a distance disparity e.g. in centimeters as shown in (1) and (2).



$$\alpha = 2 \times \text{atan}\left(\frac{DD}{2 \times VD}\right)$$

Figure 1. Disparity calculation in arcmin based on different disparities in number of pixels on display

$$w = W_{\text{cm}} / W_{\text{pixels}} \quad (1)$$

where  $W_{\text{cm}}$  is the display width in cm and  $W_{\text{pixels}}$  is the display width in pixels. Hence,  $w$  presents one pixel width in cm.

$$DD = w \times PD \quad (2)$$

where  $DD$  is the distance disparity and  $PD$  is the disparity in number of pixels

Considering the viewing distance ( $VD$ ), the disparity in arcmin can be calculated for different objects in the scene using (3). This is depicted in Figure 1.

$$D_{\text{Arcmin}} = 2 \times \text{atan}\left(\frac{DD}{2 \times VD}\right) \quad (3)$$

where  $D_{\text{Arcmin}}$  is the disparity in arcmin and  $\text{atan}$  calculates the Arc Tangent in arcmin.

Pastoor in [17] assessed the viewing comfort when watching a series of stereoscopic images with disparities ranging from 0 to 140 arcmin. The results show that disparities up to 35 arcmin do not cause any discomfort while disparities above 70 arcmin should be avoided.

#### 2.4. 3D video quality

Considering asymmetric stereoscopic video, artifacts causing contradictory depth cues are sent to each eye. Similarly to asymmetric video encoding which results in the masking of the artifacts of the worst view, the risk is to suppress the stereopsis because there might be no correspondences between the left and right views.

Even though it has been shown that image quality is important for visual comfort, it is not the only factor for great 3D visual experience. New concepts such as depth perception and presence i.e. the feeling of being there have to be considered too. These concepts are extensively studied in [25], [26], and [27].

One annoying artifact while observing 3D content with glasses is crosstalk [28]. It is perceived as shadow or double contours (ghosting artifact) due to imperfect optical

separation between the left and the right images by filters of passive glasses or slight imperfection in synchronization between shutters in active glasses and the displayed left and right views [29]. This will cause perception of opposite view by each eye causing the ghosting artifact while it should have been blocked by the viewing glasses. Crosstalk has been mentioned as one of the main disturbing perceptual display related factors for 3D viewers [30]. The ghosting artifact is most visible when watching a stereoscopic video on a 3D display without glasses (2D presentation), since both left and right views are visible to both eyes. Hence, the subjective quality of stereoscopic video in 2D presentation is not acceptable due to this artifact as depicted in Figure 2.

#### 2.5. Depth-enhanced multiview video coding

Multiview autostereoscopic displays (ASDs) require many high-quality views to be available at the decoder/display side prior to displaying. Due to the natural limitations of content production and content distribution technologies, there is no way that a large number of views can be delivered to users with existing video compression standards. Moreover, due to differing subjective preferences on the amount of depth in 3D displaying as well as different 3D displays and viewing environments, it is desirable to enable depth or disparity of the content in the decoder/display side. Therefore, the Moving Picture Experts Group (MPEG) issued a Call for Proposals for 3D video coding (hereafter referred to as the 3DV CfP) [31] for a new standard which enables rendering of a selectable number of views without increasing the required bitrate. The work initiated by MPEG has been continued in the 3D video coding standardization in the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) [32] and aims at enabling a variety of display types and



Figure 2. Subjective quality of stereoscopic video without glasses

preferences including varying camera separation to adjust the depth perception.

In ASD and other 3D display applications many views should be available at the decoder side. A multiview video plus depth (MVD) format [33], where each video data pixel is associated with a corresponding depth map value, allows reducing the input data for the 3DV systems significantly, since most of the views can be rendered from the available decoded views and depth maps using a depth-image-based rendering (DIBR) [34] algorithm. Such a scenario ensures the availability of a sufficient number of views for display where different disparities based on the targeted application can be achieved. Hence, as proposed by the 3DV CfP, a 3-view MVD coding scenario is suitable for creation of a wide range of required views for multiview ASD rendering while a suitable pair of synthesized views can also be used for rendering on a stereoscopic display.

### 3. PROPOSED RENDERING ALGORITHM

In this section, a set of adaptation methods, taking advantage of the binocular suppression theory and achieving a tradeoff between stereoscopic viewing with glasses and single-view viewing without glasses, are introduced. In these adaptation methods, one view is chosen as the dominant view while the other view will be the non-dominant view. The aim of the methods is to let the dominant view be perceived clearly and the ghosting effect caused by the non-dominant view to be close to imperceptible when viewing without glasses, while the perceived quality in viewing with glasses is only slightly degraded. The adaptation processes the non-dominant view and the disparity of the stereo pair with three methods. The selection of these methods was based on the previous conclusions in the literature showing that none of the methods is expected to affect the subjective quality of stereoscopic video considerably. In the first step, disparity is selected in agreement with [5], [17] and without sacrificing the depth perception in stereoscopic 3D presentation. Following this, the non-dominant view is low pass filtered, as it is shown in [6], [7], [13], and [14] that this does not affect the 3D perceived subjective quality. In the final step, a contrast adjustment algorithm is applied on the non-dominant view in favor of better quality in presentation without glasses. It has been confirmed in [18] that contrast adjustment of one view does not decrease the visual quality of stereoscopic video noticeably while watched with glasses. Figure 3 depicts the block diagram of the rendering process. As can be seen from Figure 3, the proposed processing takes place after decoding the stereoscopic video content and could be implemented in a television set or a display capable of stereoscopic rendering. All processing steps can be made adjustable, so that the viewers can be given the option of controlling the strength or the amount of processing. In the following sub-sections, each of the three processing steps is described in more details.

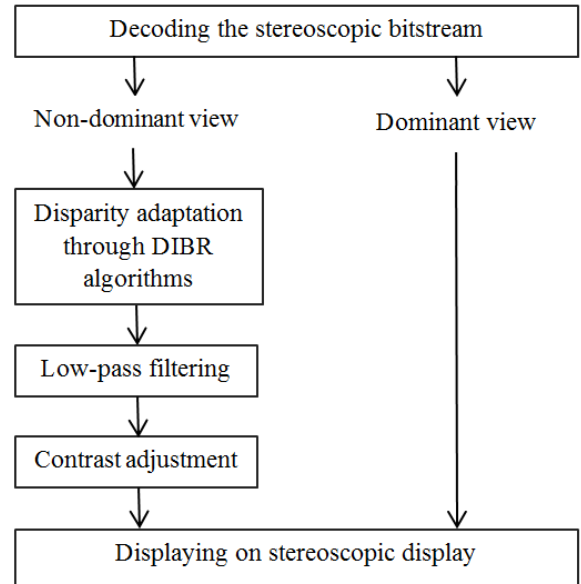


Figure 3. Block diagram of the rendering process

#### 3.1. Disparity selection

It is important to control the disparity between the views in a stereoscopic presentation of 3D content in such a manner that the content is comfortable to view while a desired depth perception is also obtained. Clearly, while increasing the distance between left and right views, the ghosting artifact in 2D presentation of stereoscopic video increases and thus, more annoying subjective quality is expected when the content is viewed without viewing glasses. On the other hand, if the small disparity between views is chosen, the depth perception in 3D presentation decreases.

Disparity selection between the views is initially determined at the time of generating the content, for example through the camera baseline separation and the distance from the camera to the filmed objects. Disparity selection at the rendering device is enabled if a depth-

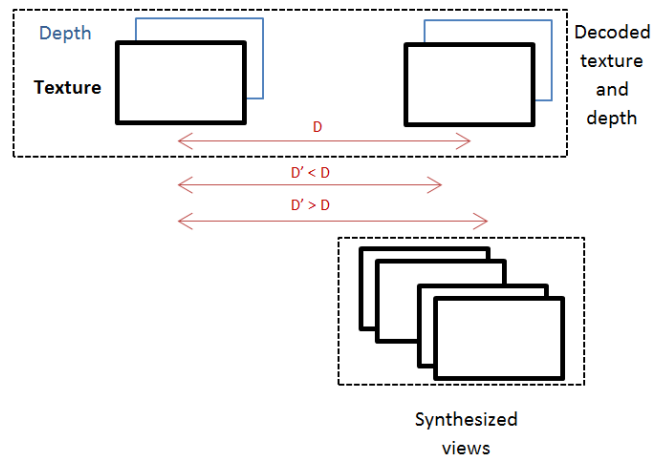


Figure 4. Enabling disparity selection through view synthesis process where  $D'$  represents the view separation achieved by view synthesis process compared to view separation of decoded views ( $D$ )

enhanced multiview video coding is used as a distribution format or if the rendering device is capable of a disparity or depth estimation from decoded stereo pairs. Consequently, by means of DIBR algorithms, a view at a desired location can be synthesized. Considering the selected disparity and hence, the estimated view separation, a combination of one coded view and one synthesized view can be exploited to create the displayed stereoscopic video. This is illustrated in Figure 4.

The proposed adaptation methods presented in the next two sub-sections aims at rendering the non-dominant view as invisible as possible in the presentation of stereoscopic video without glasses. Nevertheless, having a smaller disparity still provides a smoother subjective quality for a 2D presentation of the content.

### 3.2. Low pass filtering

Low-pass filtering decreases the number of high frequency components (HFCs) in the non-dominant view by removing some details. Hence, in the created asymmetric stereoscopic video, the non-dominant view will be blurred compared to the dominant view. This will favor better 2D presentation of the stereo pair, as the dominant view will be sharper compared to the blurred non-dominant view and therefore it will be better perceived by HVS. Yet, as verified extensively in previous studies [4], [6], [7], and [14] asymmetric stereoscopic video where one view has been low pass filtered provides similar subjective quality and depth perception to those of stereoscopic video where both views have the same high quality.

In our experiments, the applied low-pass filter (LPF) was a 2D circular averaging filter (pillbox) within a square matrix having  $2 \times \text{radius} + 1$  elements in each row and column, as it resulted in a better subjective performance compared to a few other tested LPFs. The equation used for this filter is MATLAB implementation of a simple pillbox filter presented in [35]. In general, any LPF can also be selected for example on the basis of memory access and complexity constraints. The level of HFC reduction depends on the radius defined for the filter such that increasing the radius results in more reduction of HFCs. The 2D matrix presenting the LPF coefficients of the used LPF for radius 6 is depicted in (4).

$$f = 10^{-4} \times \begin{bmatrix} 0 & 0 & 0 & 0 & 13 & 36 & 44 & 36 & 13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 61 & 88 & 88 & 88 & 88 & 88 & 61 & 8 & 0 & 0 \\ 0 & 8 & 76 & 88 & 88 & 88 & 88 & 88 & 88 & 76 & 8 & 0 & 0 \\ 0 & 61 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 61 & 0 & 0 \\ 13 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 13 & 0 \\ 36 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 36 & 0 \\ 44 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 44 & 0 \\ 36 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 36 & 0 \\ 13 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 13 & 0 \\ 0 & 61 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 61 & 0 & 0 & 0 \\ 0 & 8 & 76 & 88 & 88 & 88 & 88 & 88 & 88 & 76 & 8 & 0 & 0 \\ 0 & 0 & 8 & 61 & 88 & 88 & 88 & 88 & 88 & 61 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 13 & 36 & 44 & 36 & 13 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$

### 3.3. Contrast adjustment

The response of HVS depends much more on the relation of luminance local variations compared to the surrounding

values than absolute luminance. Contrast is a measure for this relative variation of luminance. In the visual perception of different scenes, contrast is determined by the difference in color and brightness of each object and other objects in the same viewing field. Hence, contrast adjustment is related to brightness and color settings i.e. how the luminance and chrominance differ and change.

The approach utilized in this experiment is to decrease the contrast of luma and chroma components of the non-dominant view while keeping the contrast of the dominant view unchanged. The contrast decrease of the non-dominant view will help the 2D presentation of the stereoscopic view that has more similarity to the dominant view while the stereoscopic presentation is not influenced considerably.

The contrast adjustment of an image can be done in various ways. We follow the same algorithm as used for the weighted prediction mode of the Advanced Video Coding (H.264/AVC) standard [36], that is:

$$O = \text{round} \left( \frac{i \times w}{2^d} \right) = (i \times w + 2^{d-1}) \gg d \quad (5)$$

where:

$O$  is the adjusted luma or chroma contrast value

$\text{round}$  is a function returning the closest integer

$i$  is the input sample value

$w$  and  $d$  are the parameters utilized to create the adjustment weight

$\gg$  is a bit shift operation to the right

## 4. TEST SETUP

The performed tests targeted at verifying that the proposed method has potential to tackle the presented problem satisfactorily, i.e. that the same stereoscopic 3D content can be viewed with viewing glasses with acceptable 3D quality and depth perception and without viewing glasses with acceptable 2D quality and a tolerable level of ghosting artifacts. Furthermore, the performed tests aimed at discovering how to tune the processing steps of the proposed algorithm optimally, i.e. which are good trade-offs for the three processing components, disparity selection, low-pass filtering and contrast adjustment. As no objective video quality metrics are applicable to the presented problem as far as the authors are aware of, a large-scale subjective assessment was performed with four sequences: Poznan Hall2, Poznan Street [37], Ghost Town Fly (GT Fly), Undo Dancer, which were used in the 3DV CFP [31]. For GT Fly and Undo Dancer sequences, 500 frames were used while 250 and 200 frames were used for Street and Hall2, respectively. No encoding was applied to the

TABLE I. INPUT VIEWS AND CAMERA DISTANCES FOR SMALL AND BIG CAMERA SEPARATIONS

Sequence	Left view-Right view , (Camera separation in cm)	
	Small disparity	Big disparity
Poznan Hall2	7-6.5 , (6.87)	7-6 , (13.75)
Poznan Street	5-4.5 , (6.87)	5-4 , (13.75)
GT Fly	3-1 , (4)	5-1 , (8)
Undo Dancer	1-3 , (4)	1-5 , (8)

sequences. The frame rate was fixed to 25 Hz for all sequences. Each sequence was evaluated at two different disparities or camera separations, referred to as small and big disparity subsequently. The camera separation of the big disparity is the same as those introduced in MPEG 3DV CfP for the 3-view coding scenario and can be considered to represent a typical disparity for stereoscopic viewing, while in the small disparity scheme the camera separation distance is halved. The input views and the relative camera separation distances used in the experiments, for both small and large disparity stereoscopic sequences, are shown in Table 1.

#### 4.1. Preparation of Test Stimuli

To prepare test material, the three adaptation methods presented in section 3 were used and various test cases based on different combinations of adaptation methods were created. In the experiments, we tested contrast reduction to 50% and 75% of the original values for different combinations by fixing the value of  $d$  to 4 and setting the value of  $w$  equal to 8 and 12, respectively, in equation (5). Moreover, all non-dominant views for different schemes except Original 2D were low pass filtered using the circular averaging filter with radius equal to 6 as presented in equation (4).

Two different disparities between the left and the right views were selected for different sequences. In the test sequences the disparity was always positive i.e. the objects are always behind the display level. Disparity selection was limited so that the results were in agreement with previous findings in the literature to prevent eye strain due to excessive disparities.

Disparity can be calculated from depth map by converting it to disparity. Table 2 presents the average and the maximum disparities for each sequence. Moreover, Table 1 presents the selected views and corresponding camera separations for different disparities of the sequences. For Poznan Hall2 and Poznan Street sequences, views 6.5 and 4.5, respectively, were synthesized from the original texture and depth views using the MPEG View Synthesis Reference Software (VSRS) version 3.5 [38]. The subjective quality of synthesized views was comparable to that of the original views. Moreover, since the synthesized artifacts were subjectively negligible, we assume that the synthesizing process did not affect the subjective ratings.

Combining the above-mentioned tested parameters, the following seven test cases were prepared and subjectively assessed. The combinations for each scheme are presented in the format of (disparity, contrast) where for disparity the

TABLE 2. DISPARITIES FOR SMALL AND BIG CAMERA SEPARATION

Sequence	Average disparity (Maximum disparity) in arcmin	
	Small disparity	Big disparity
Poznan Hall2	18.6(22.2)	37.2(44.3)
Poznan Street	19.3(23.6)	38.6(47.2)
GT Fly	12.1(42.2)	24.3(84.3)
Undo Dancer	13.6(22.2)	27.2(47.2)

TABLE 3. DIFFERENT SCHEMES AND THEIR CHARACTERISTICS

Scheme	Disparity	Contrast adjustment
<b>O</b> → (Original 2D)	<i>0</i>	100%
<b>S1</b>	<i>Small</i>	100%
<b>S2</b>	<i>Small</i>	75%
<b>S3</b> → (Best 2D quality)	<i>Small</i>	50%
<b>B1</b> → (Best 3D quality)	<i>Big</i>	100%
<b>B2</b>	<i>Big</i>	75%
<b>B3</b>	<i>Big</i>	50%

values *0*, *Small*, *Big* refer to *0* disparity (identical left and right views), *Small* disparity, and *Big* disparity, respectively. For contrast the values  $X\%$  present the contrast ratio of the non-dominant view relative to the dominant view. Seven different test schemes, as presented in Table 3, were used in the subjective tests.

#### 4.2. Test Procedure and Subjects

Subjective viewing was conducted according to the conditions suggested in MPEG 3DV CfP. The polarized 46'' Vuon E465SV 3D display manufactured by Hyundai was used. The display has a total resolution of 1920×1200 pixels and a resolution of 1920×600 per view when used in the stereoscopic mode was used for displaying the test material. The viewing distance was equal to 4 times the displayed image height (2.29m).

Subjective quality assessment was done according to the Double Stimulus Impairment Scale (DSIS) method [39] with a discrete unlabeled quality scale from 0 to 10 for quality assessment. The test was divided into two sessions where in the first session, subjects assessed the subjective quality of video clips with glasses and in the second session, the test was performed without glasses. Two questions for each session of the test were considered and the subjects wrote their ratings after each clip was played. These questions are presented in Table 4. Each question is associated with its short term for simplicity in reporting the results. Prior to each test, subjects were familiarized with the test task, the test sequences and the variation in the quality to be expected in the actual tests. The subjects were instructed that 0 stands for the lowest quality and 10 for the highest.

Subjective viewing was conducted with 20 subjects, (16 males, 4 females), aged between 21-31 years (mean: 24.2). All subjects passed the test for stereovision prior to the actual test. Moreover, they were all considered naïve as they did not work or study in fields related to information technology, television or video processing. To prevent subjects from getting exhausted during the evaluation sessions, the duration of the test was limited to 45 minutes.

## 5. RESULTS AND DISCUSSION

In this section we present the results of the conducted subjective tests and an analysis of the statistics of the quantitative viewing experience ratings.

Figure 5 shows the subjective viewing experience ratings with 95% confidence interval (CI) for all sequences. The results are provided for four questions that subjects

TABLE 5. FLAG TABLE PRESENTING SIGNIFICANT DIFFERENCES FOR DIFFERENT TEST SCHEMES PRESENTED IN TABLE 3

FLAGS -1, 0, AND 1 PRESENT SIGNIFICANTLY LOWER, SIMILAR, AND SIGNIFICANTLY HIGHER QUALITY COMPARED TO OTHER SCHEMES, RESPECTIVELY

	Flags	Test scheme combinations					
		S1	S2	S3	B1	B2	B3
Dancer	-1	2	<b>1</b>	0	3	2	3
	0	17	<b>17</b>	16	16	18	14
	1	1	<b>2</b>	4	1	0	3
GT Fly	-1	4	<b>1</b>	1	5	2	1
	0	16	<b>17</b>	13	13	17	16
	1	0	<b>2</b>	6	2	1	3
Street	-1	4	<b>2</b>	3	7	4	4
	0	13	<b>13</b>	9	10	15	12
	1	3	<b>5</b>	8	3	1	4
Hall2	-1	4	<b>3</b>	6	4	0	4
	0	12	<b>14</b>	12	12	14	14
	1	4	<b>3</b>	2	4	6	2
<b>Sum for all sequences</b>	-1	14	<b>7</b>	10	19	8	12
	0	58	<b>61</b>	50	51	64	56
	1	8	<b>12</b>	20	10	8	12

were asked during the test sessions (see sub-section 4.2). The naming introduced for the different schemes in sub-section 4.1 is used in the figures for simplicity. Subjective ratings show that scheme O achieved the highest value in 2D evaluation (i.e. the session where viewing took place without glasses) and in the general quality of 3D presentation. However, because depth perception was rated the smallest in this scheme, it cannot be considered as a competitor for an acceptable trade-off for simultaneous 2D and 3D perception. Hence, it was excluded from the analysis presented next. For the other tested schemes, the following general trend was observed. In both small and big disparities, while decreasing the contrast ratio of the non-dominant view, the ratings of the 2D evaluation session increase and at the same time the 3D evaluation ratings decrease. This was expected as reducing the contrast of the non-dominant view targets ideal 2D subjective quality while compromising the 3D perception. Moreover, in all sequences, the ghosting effect in the 2D presentation of stereoscopic video clips without any contrast adjustment annoyed subjects more in the big disparity scheme when compared to the small disparity schemes. Considering the large amount of viewing experience ratings, it is hard to make many logical conclusions based on Figure 5. Hence, significant differences between the schemes were further

analyzed using statistical analysis as presented in the paragraphs below.

The Wilcoxon's signed-rank test [40] was used as the data did not reach normal distribution (Kolmogorov-Smirnov:  $p < 0.05$ ). Wilcoxon's test is used to measure differences between two related and ordinal data sets [41]. A significance level of  $p < 0.05$  was used in the analysis.

The following conclusions were obtained with this statistical significance analysis mentioned above. In the analysis, we compared pairwise the ratings of each two test case combinations resulting in fifteen flags per question and per sequence, indicating whether the subjective quality between different test cases have any statistically significant difference. Considering four sequences, four questions per sequence, and fifteen two-sided pairwise comparisons per question, we obtained  $4 \times 4 \times 15 \times 2 = 480$  flags. Table 5 reports a summary of the distribution of these flags. Each cell presents the total number of flags from different questions where -1, 0, and 1 present significantly lower, similar, and significantly higher quality compared to other schemes, respectively. From this Table it is clear that only S2 provides similar or better subjective results for all sequences while other schemes have a lower performance at least in one sequence. Hence, the combination used in S2 seems to be a well-designed potential candidate for simultaneous 2D and 3D presentation. Moreover, Table 5 reports the cumulative flag counts over all sequences. It can be observed that the cumulative counts for S3 are comparable or better than those for S2. However, by studying the performance of S3 for individual sequences, it can be observed that the performance of S3 for Hall2 is inferior to the results obtained with S2. To analyze the subjective performance of each test scheme combination for 2D and 3D viewing separately, similar flag tables as the one presented in Table 5 are presented in Table 6, reporting results for 2D and 3D viewing experiments separately. Considering the two summaries provided in Table 6, S2 is the only test scheme for which the number of test cases where its performance was statistically superior to the another test scheme (flag value equal to 1) was greater than the number of test cases where its performance was statistically inferior to another test scheme (flag value equal to -1) in both 2D and 3D viewing experiments.

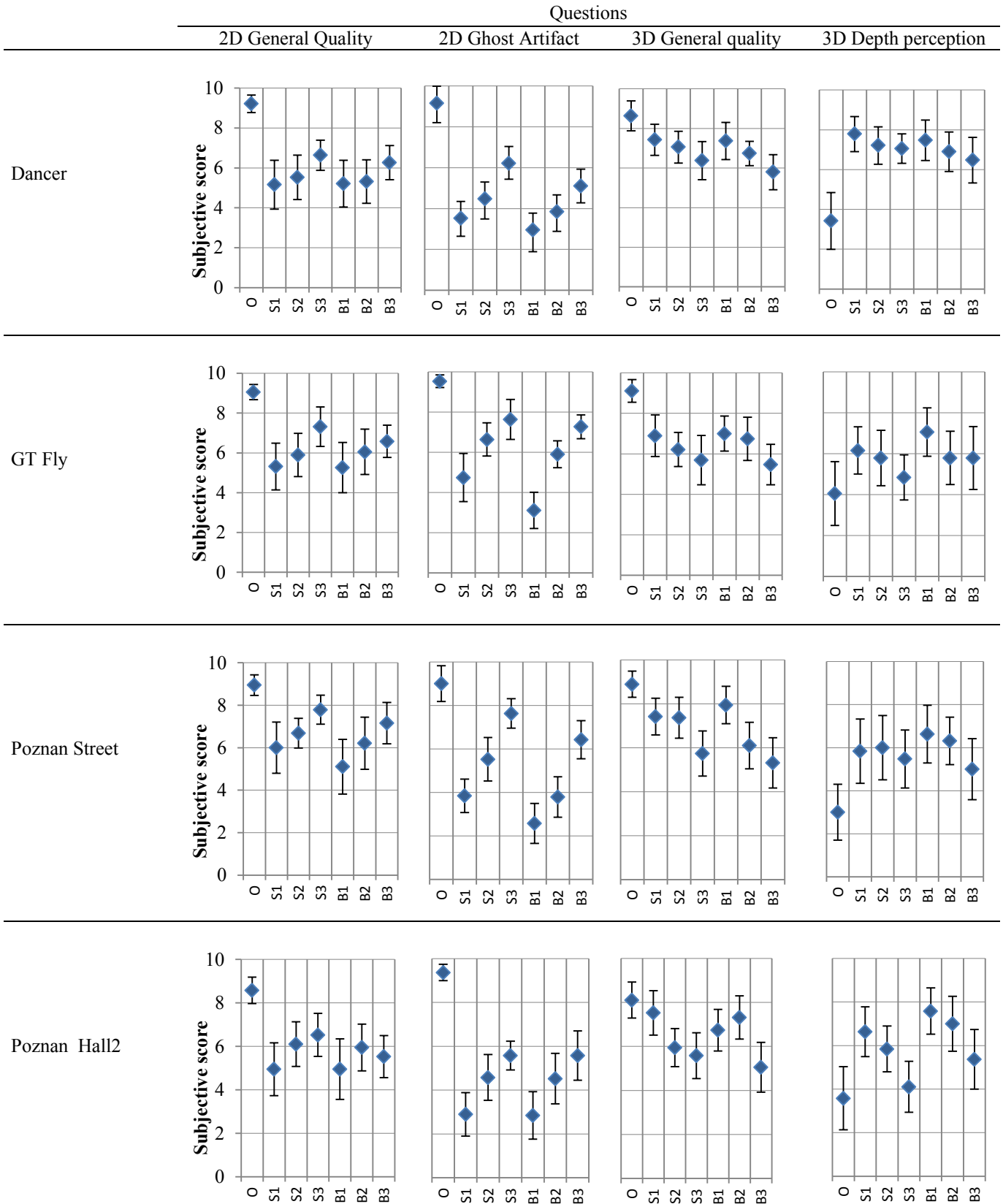


Figure 5. Viewing experience ratings with 95% confidence interval. The schemes are named according to Table 3.

TABLE 6. FLAG TABLE PRESENTING SIGNIFICANT DIFFERENCES FOR DIFFERENT TEST SCHEMES PRESENTED IN TABLE 3 FOR (A) 2D AND (B) 3D EXPERIMENTS

FLAGS -1, 0, AND 1 PRESENT SIGNIFICANTLY LOWER, SIMILAR, AND SIGNIFICANTLY HIGHER QUALITY COMPARED TO OTHER SCHEMES, RESPECTIVELY

		(A)					
		Test scheme combinations					
	Flags	S1	S2	S3	B1	B2	B3
Dancer	-1	2	<b>1</b>	0	3	2	0
	0	8	<b>8</b>	6	7	8	7
	1	0	<b>1</b>	4	0	0	3
GT Fly	-1	4	<b>1</b>	0	5	2	0
	0	6	7	4	5	7	7
	1	0	<b>2</b>	6	0	1	3
Street	-1	4	<b>2</b>	0	7	3	1
	0	5	<b>5</b>	2	3	6	5
	1	1	<b>3</b>	8	0	1	4
Hall2	-1	4	<b>0</b>	0	4	0	0
	0	6	<b>8</b>	8	6	8	8
	1	0	<b>2</b>	2	0	2	2
<b>Sum for all sequences</b>	-1	14	<b>4</b>	0	19	7	1
	0	25	<b>28</b>	20	21	29	27
	1	1	<b>8</b>	20	0	4	12

		(B)					
		Test scheme combinations					
	Flags	S1	S2	S3	B1	B2	B3
Dancer	-1	0	<b>0</b>	0	0	0	3
	0	9	<b>9</b>	10	9	10	7
	1	1	<b>1</b>	0	1	0	0
GT Fly	-1	0	<b>0</b>	1	0	0	1
	0	10	<b>10</b>	9	8	10	9
	1	0	<b>0</b>	0	2	0	0
Street	-1	0	<b>0</b>	3	0	1	3
	0	8	<b>8</b>	7	7	9	7
	1	2	<b>2</b>	0	3	0	0
Hall2	-1	0	<b>3</b>	6	0	0	4
	0	6	<b>6</b>	4	6	6	6
	1	4	<b>1</b>	0	4	4	0
<b>Sum for all sequences</b>	-1	0	<b>3</b>	10	0	1	11
	0	33	<b>33</b>	30	30	35	29
	1	7	<b>4</b>	0	10	4	0

The conclusion that S2 provides the most acceptable trade-off for simultaneous 2D and 3D viewing is in agreement with previous findings on contrast asymmetry in [18], where the contrast difference limit between the left and the right views was found to be equal to or less than 25% to provide equal viewing comfort. Moreover, considering camera separations presented in Table 2, the perceived disparity for all sequences was aligned with the results presented in [17], [23], and [24], where the limit for the maximum disparity between the left and right views was found to be 70 arcmin. Only the maximum disparity of the big camera separation for GT Fly is above this limit. This big disparity happens for 0.06 seconds in the 20 second sequence (3 frames in 500 frames). Figure 6 depicts a sample frame from a 2D presentation of a stereoscopic video from scheme S2 and the corresponding stereoscopic video frame with equal disparity and without any LPF or contrast adjustment applied.

After the test, the participants were asked whether they experienced any fatigue or eye strain during and/or after the test. Subjects seemed quite comfortable and there were no complaints regarding the 3D content and the asymmetric nature of the stereoscopic video clips. However, five subjects complained that sometimes it was difficult to distinguish the differences between the observed clips.

## 6. CONCLUSION

Stereoscopic video provides 3D perception by presenting slightly different views for each eye. Ghosting artifacts make it almost intolerable to watch the content without glasses for both active and passive glasses/displays. In this paper we tackled the problem of viewing 3D content simultaneously with and without viewing glasses by proposing a technique which makes it quite acceptable to watch stereoscopic content without glasses while the 3D perception is not sacrificed much. In the proposed approach, one dominant view is selected and then the non-dominant view is adjusted through disparity selection, contrast adjustment, and low-pass-filtering. These steps increase the similarity of the non-dominant view to the dominant view.

The performance of the proposed technique was assessed through extensive subjective tests. The statistical analysis of scores showed that combination of a disparity smaller than what is conventionally used for stereoscopic video along with low-pass-filtering the non-dominant view and decreasing its contrast to 75% provides the best trade-off between 3D and 2D perception of a stereoscopic 3D content. This is a new topic introduced in 3D research field and as a future plan we intend to do more research on other potential approaches to be used in the process.





Hall2



Street



Dancer

(a) (b)

Figure 6. 2D presentation of stereoscopic video combinations from (a) Original scheme and (b) Selected scheme i.e. S2

#### ACKNOWLEDGEMENT

The authors thank Prof. M. Domański, et al. for providing Poznan sequences and their camera parameters [37].

#### REFERENCES

- [1] C. Wheatstone, "On some remarkable, and hitherto unobserved, phenomena of binocular vision," *Philos. Trans. R. Soc. Lond.*, 1838.
- [2] H. von Helmholtz, "Handbuch der physiologischen Optik," Leopold Voss 1866, (English ed. 1962, Dover New York).
- [3] H. Asher, "Suppression theory of binocular vision," *The British Journal of Ophthalmology*, vol. 37, no. 1, pp. 37-49, 1953.
- [4] B. Julesz, "Foundations of Cyclopean Perception," University of Chicago Press, Chicago, IL, USA, 1971.
- [5] P. Seuntjens, L. Meesters, and W. IJsselstein, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Appl. Perception (TAP)*, vol. 3, pp. 95-109, 2006.
- [6] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video," *Proc. of 3DTV Conference*, June 2010.
- [7] L. B. Stelmach, W. J. Tam, D. V. Meegan, A. Vincent, and P. Corriveau, "Human perception of mismatched stereoscopic 3D inputs," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Sep. 2000.
- [8] P. Aflaki, M. M. Hannuksela, J. Hakala, J. Häkkinen, and M. Gabbouj, "Joint adaptation of spatial resolution and sample value quantization for asymmetric stereoscopic video compression: a subjective study," *Proc. of International Symposium on Image and Signal Processing and Analysis*, Sep. 2011.
- [9] F. Shao, G. Jiang, X. Wang, M. Yu, and K. Chen, "Stereoscopic video coding with asymmetric luminance and chrominance qualities," *IEEE Trans. on Consumer Electronics*, vol.56, no.4, pp.2460-2468, Nov. 2010.
- [10] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Sep. 2010.
- [11] Z. Zivkovic, E. Bellers, "Efficient and robust detection of users wearing stereoscopic glasses for automatic 2D/3D display mode selection", *International Conference on Consumer Electronics (ICCE)*, pp. 682-683, Jan, 2012
- [12] L. M. Wilcox, J. H. Elder, and R. F. Hess, "The effects of blur and size on monocular and stereoscopic localization," *Vision Research*, vol. 40, pp. 3575-3584, 2000.
- [13] G. Papalba, I. Cipane, and M. Ozolinsh, "Stereo vision studies by disbalanced images," *Advanced Optical Devices, Technologies, and Medical Applications*, *Proc. of SPIE* vol. 5123, pp. 323-329, 2003.
- [14] L. Stelmach, W. J. Tam; D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 10, No. 2, pp. 188-193, March 2000.
- [15] M. G. Perkins, "Data compression of stereopairs," *IEEE Trans. on Communications*, vol. 40, no. 4, pp. 684-696, Apr. 1992.
- [16] A. Aksay, C. Bilen, and G. Bozdagi Akar, "Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images," *European Signal Processing Conference, EUSIPCO*, Sep. 2005.
- [17] S. Pastoor, "Human factors of 3D imaging: results of recent research at Heinrich-Hertz-Institut Berlin," *Proceedings of the International Display Workshop (Asia Display)*, 1995.
- [18] F. L. Kooi and A. Toet, "Visual Comfort of binocular and 3D displays," in *Displays*, Elsevier, vol. 25, pp. 99-108, 2004.
- [19] R. Suryakumar, "Study of the Dynamic Interactions between Vergence and Accommodation," University of Waterloo, 2005.
- [20] P. A. Howarth, "Empirical studies of accommodation, convergence, and HMD use," *Proceedings of the Hoso-Bunka Foundation Symposium: The Human Factors in 3-D Imaging*, 1996.

- [21] A. Coltekin, "Foveation for 3D visualisation and stereo imaging," Ph.D. thesis, ISBN 951-22-8017-5, Helsinki University of Technology, 2006.
- [22] D. Qin, M. Takamatsu, Y. Nakashima, "Measurement for the Panum's Fusional Area in Retinal Fovea Using a Three-Dimension Display Device," *Journal of Light & Visual Environment*, pp. 126-131, Vol. 28, Issue 3, 2004.
- [23] F. Speranza, W. J. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," *Proc. Stereoscopic Displays and Virtual Reality Syst. XIII*, vol. 6055, pp. 60550B-1–60550B-9, 2006.
- [24] M. Wopking, "Viewing comfort with stereoscopic pictures: an experimental study on the subjective effects of disparity magnitude and depth of focus," *Journal of the Society for Information Display* 3: 101-103, 1995.
- [25] W. A. IJsselsteijn, H. de Ridder, J. Freeman, and S. E. Avons, "Presence: concept, determinants, and measurement," *Proc. of the SPIE*, Vol. 3959, pp. 520-529, 2000.
- [26] W. A. IJsselsteijn, H. de Ridder, J. Freeman, S. E. Avons., and D. Bouwhuis, "Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence," *Presence-Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 298-311, 2001.
- [27] P. Seuntjens, *Visual Experience of 3D TV*, Ph.D. thesis, 2006.
- [28] J. Konrad and M. Halle, "3-D displays and signal processing: An answer to 3-D ills?," *IEEE Signal Process. Mag.*, vol. 24, pp. 97-111, Nov. 2007.
- [29] D. Strohmeier, S. Jumisko-Pyykkö, U. Reiter, "Profiling experienced quality factors of audiovisual 3D perception", *Proc. of the Int. Workshop on Quality of Multimedia Experience*, pp. 70-75, June 2010
- [30] W. A. IJsselsteijn, P. H. J. Seuntjens, and L. M. J. Meesters, "Human Factors of 3D Displays," *3D Video communication, algorithms, concepts, and real-time systems in human-centred communication*, John Wiley & Sons, Ltd.:219-234, 2005.
- [31] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, March 2011. [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3dv-cfp.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3dv-cfp.zip)
- [32] <http://phenix.int-evry.fr/jct3v/>
- [33] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, vol. 1, pp. 201-204, Oct. 2007.
- [34] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in *Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 93–104, Jan. 2004.
- [35] Xin Wang, Baofeng Tian, Chao Liang, "Blind image quality assessment for measuring image blur," *IEEE, 2008 Congress on Image and Signal Processing*, pp. 467-470, May, 2008.
- [36] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," Jan. 2012.
- [37] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters," *MPEG 2009/M17050*, Oct. 2009.
- [38] "View synthesis software manual," *MPEG ISO/IEC JTC1/SC29/WG11*, Sept. 2009.
- [39] ITU-R Rec. BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, 2002.
- [40] F. Wilcoxon. *Individual comparisons by ranking methods*. *Biometrics*, 1:80–83, 1945.
- [41] H. Cooligan "Research methods and statistics in psychology," (4th ed.). London: Arrowsmith., 2004.

[P3] **P. Aflaki**, M. M. Hannuksela, and M. Gabbouj; “Subjective quality assessment of asymmetric stereoscopic 3-D video,” Springer Journal of Signal, Image and Video Processing, March, 2013.

© Springer, 2013, Reprinted with permission.

# Subjective quality assessment of asymmetric stereoscopic 3D video

Payman Aflaki · Miska M. Hannuksela ·  
Moncef Gabbouj

Received: 15 May 2012 / Revised: 12 December 2012 / Accepted: 22 February 2013  
© The Author(s) 2013. This article is published with open access at Springerlink.com

**Abstract** In asymmetric stereoscopic video compression, the views are coded with different qualities. According to the binocular suppression theory, the perceived quality is closer to that of the higher-fidelity view. Hence, a higher compression ratio is potentially achieved through asymmetric coding. Furthermore, when mixed-resolution coding is applied, the complexity of the coding and decoding is reduced. In this paper, we study whether asymmetric stereoscopic video coding achieves the mentioned claimed benefits. Two sets of systematic subjective quality evaluation experiments are presented in the paper. In the first set of the experiments, we analyze the extent of downsampling for the lower-resolution view in mixed-resolution stereoscopic videos. We show that the lower-resolution view becomes dominant in the subjective quality rating at a certain downsampling ratio, and this is dependent on the sequence, the angular resolution, and the angular width. In the second set of the experiments, we compare symmetric stereoscopic video coding, quality-asymmetric stereoscopic video coding, and mixed-resolution coding subjectively. We show that in many cases, mixed-resolution coding achieves a similar subjective quality to that

of symmetric stereoscopic video coding, while the computational complexity is significantly reduced.

**Keywords** Mixed resolution · Asymmetric stereoscopic · Stereoscopic 3D video · Subjective quality

## 1 Introduction

Stereoscopic video compression has gained importance during the recent years thanks to the recent advances in display technology. In many stereoscopic 3D video services and applications, the challenge is that the available bitrate or storage space is similar to that for monoscopic video, while the perceived temporal and spatial quality should also be similar to those for monoscopic video. Recent advances in video compression have alleviated the mentioned challenge to some extent. For example, the inter-view prediction enabled by the Multiview Video Coding (MVC) [1] annex of the widely used Advanced Video Coding (H.264/AVC) standard [2] has been shown to improve compression efficiency significantly compared to independent coding of the views. As an example, Merkle et al. [3] reported gains up to 3.2 dB and an average gain of 1.5 dB in terms of average luma peak signal-to-noise ratio (PSNR). However, further compression without compromising the visual quality is desirable in order to meet the bitrate and quality expectations of many applications. There are several other examples for video coding methods that aim to provide higher performance encoding to video content, for example, High Efficiency Video Coding (HEVC) [4] and a depth enhanced extension for MVC, abbreviated MVC+D, specifying encapsulation of MVC-coded texture and depth views into a single bitstream [5,6].

---

This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006–2011).

---

P. Aflaki (✉) · M. Gabbouj  
Department of Signal Processing,  
Tampere University of Technology,  
Tampere, Finland  
e-mail: payman.aflaki@tut.fi

M. Gabbouj  
e-mail: moncef.gabbouj@tut.fi

M. M. Hannuksela  
Nokia Research Center, Tampere, Finland  
e-mail: miska.hannuksela@nokia.com

Video compression is commonly achieved by removing spatial, frequency, and temporal redundancies. Different types of prediction and quantization of transform-domain prediction residuals are jointly used in many video coding standards to exploit both spatial and temporal redundancies. In addition, as coding schemes have a practical limit in the redundancy that can be removed, spatial and temporal sampling frequency as well as the bit depth of samples can be selected in such a manner that the subjective quality is degraded as little as possible.

One branch of research for obtaining compression improvement in stereoscopic video is known as asymmetric stereoscopic video coding, in which there is a quality difference between the two coded views. This is attributed to the binocular suppression theory [7]. It is assumed according to the binocular suppression theory that the HVS fuses the two images with different levels of sharpness such that the perceived quality is close to that of the sharper view [8]. This is because, in normal vision, there is some additional fusion to impulses from corresponding points of the two retinas. The correspondence of the retinal elements is completely rigid and un-changing; however, one of a pair of corresponding points tends to suppress the other and create the binocular suppression. In the next sections, we will cover several studies which have been exploiting binocular suppression in asymmetric stereoscopic video coding.

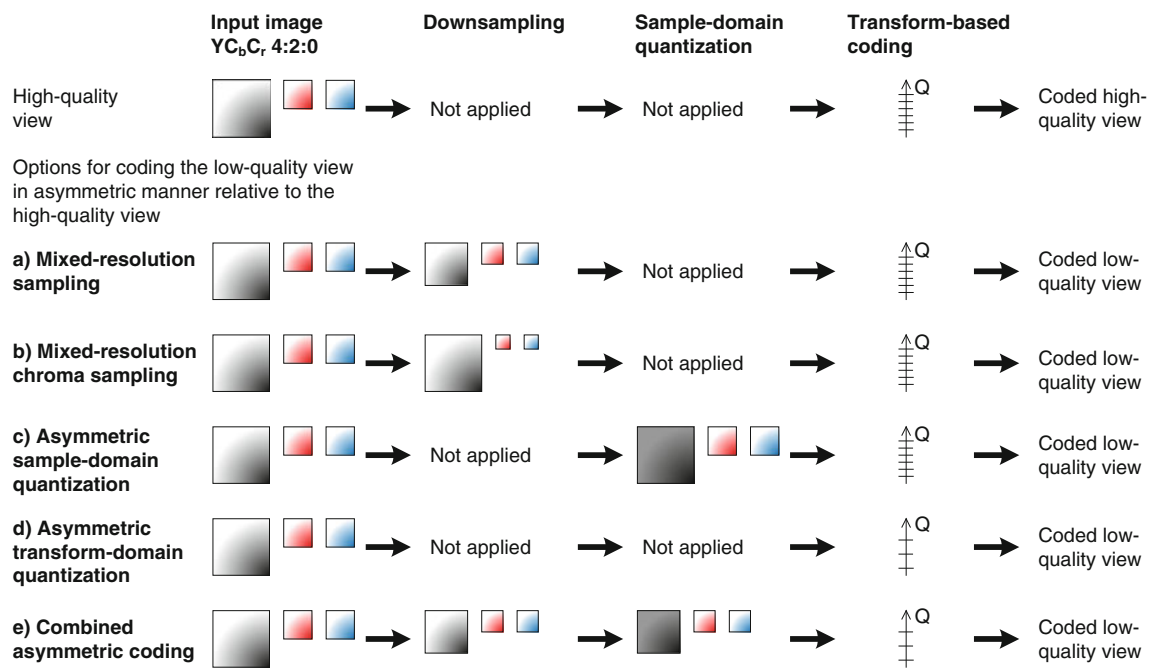
Asymmetry in quality between the two coded views can be achieved by one or more of the following methods:

- (a) Mixed-resolution (MR) stereoscopic video coding, first introduced in [9], also referred to as resolution-asymmetric stereoscopic video coding. One of the views is low-pass filtered and hence has a smaller amount of spatial details or a lower spatial resolution. Furthermore, the low-pass filtered view is usually sampled with a coarser sampling grid, that is, represented by fewer pixels.
- (b) Mixed-resolution chroma sampling [10]. The chroma pictures of one view are represented by fewer samples than the respective chroma pictures of the other view.
- (c) Asymmetric sample-domain quantization [11]. The sample values of the two views are quantized with a different step size. For example, the luma samples of one view may be represented with the range of 0–255 (i.e., 8 bits per sample), while the range may be scaled to the range of 0–159 for the second view. Thanks to fewer quantization steps, the second view can be compressed with a higher ratio compared to the first view. Different quantization step sizes may be used for luma and chroma samples. As a special case of asymmetric sample-domain quantization, one can refer to bit-depth-asymmetric stereoscopic video when the number of quantization steps in each view matches a power of two.
- (d) Asymmetric transform-domain quantization. The transform coefficients of the two views are quantized with a different step size. As a result, one of the views has a lower fidelity and may be subject to a greater amount of visible coding artifacts, such as blocking and ringing.
- (e) A combination of different encoding techniques above.

The aforementioned types of asymmetric stereoscopic video coding are illustrated in Fig. 1. The first row presents the higher quality view which is only transform-coded. The remaining rows present several encoding combinations which have been investigated to create the lower quality view using different steps, namely, downsampling, sample-domain quantization, and transform-based coding. It can be observed from the figure that downsampling or sample-domain quantization can be applied or skipped regardless of how other steps in the processing chain are applied. Likewise, the quantization step in the transform-domain coding step can be selected independently of the other steps. Thus, practical realizations of asymmetric stereoscopic video coding may use appropriate techniques for achieving asymmetry in a combined manner as illustrated in Fig. 1e. Moreover, in [12], the subjective quality of mixed temporal resolution was assessed and compared to mixed spatial resolution on two test sequences having a resolution of  $720 \times 480$ . The paper concluded that at 1/2 temporal resolution, mixed temporal resolution performed worse than mixed spatial resolution with different downsampling ratios. Due to its inferior performance, mixed temporal resolution is not considered in the subsequent parts of this paper.

This paper attempts to provide answers to two research questions: Firstly, to what extent downsampling can be applied for mixed resolution stereoscopic video? Secondly, what are the constraints which limit the preference of utilizing asymmetric coding achieved with different coding schemes compared to symmetric coding? These research questions were studied using systematic subjective testing, because no commonly acceptable objective metrics are available for approximating the perceived quality of asymmetric stereoscopic video.

The rest of this paper is organized as follows: A brief overview of the relevant literature is presented in Sect. 2. Section 3 presents a study of downsampling constraints for MR stereoscopic video. Asymmetric stereoscopic video achieved by mixed-resolution coding or asymmetric transform-domain quantization is subjectively assessed and compared to symmetric stereoscopic video coding in Sect. 4. The primary target in the study presented in Sect. 4 is to reveal whether asymmetric stereoscopic video coding outperforms symmetric stereoscopic video coding in terms of subjective quality when the same bitrate is used for both. Furthermore, the study compares the subjective quality achieved by the



**Fig. 1** Illustrative examples of different types of asymmetric stereoscopic video coding

mentioned two asymmetric stereoscopic video coding methods. Finally, conclusions are provided in Sect. 5.

## 2 Literature review

### 2.1 Uncompressed mixed-resolution stereoscopic video

The subjective impact of uncompressed MR sequences at downsampling ratios of 1/2 and 1/4 applied both horizontally and vertically was studied in [12]. A combination of a data projector and shutter glasses were used as the viewing equipment with a viewing distance equal to 4H, where H was 91.5 cm. It was found that the perceived sharpness and the subjective image quality of the MR image sequences were nearly transparent at the downsampling ratio of 1/2 along both coordinate axes but dropped slightly at the ratio of 1/4.

The study presented in [13] included a subjective evaluation for full- and mixed-resolution stereo video on a 32-inch polarization stereo display and on a 3.5-inch mobile display. One of the views in the MR sequences was downsampled to half the resolution both horizontally and vertically. The results revealed that uncompressed full-resolution (FR) sequences were preferred in 94 and 63 % of the test cases for the 32- and 3.5-inch displays, respectively. Moreover, different resolutions for the symmetric stereo video and the higher-resolution view of the MR videos were tried out, while the downsampling ratio in the MR videos was always 1/2 both horizontally and vertically. It was found that the

higher the resolution, the smaller the subjective difference is between FR and MR stereoscopic video. An equivalent result was also discovered as a function of the viewing distance by changing the distance from 1 to 3 m—the greater the viewing distance, the smaller the subjective difference becomes between FR and MR.

An obvious question related to MR stereoscopic video is whether people having a different ocular dominance perceive the quality of the same MR stereoscopic image sequence differently. However, it has been discovered in several studies, such as [14] and [15], that subjective ratings of MR image sequences are not statistically impacted by eye dominance.

In this paper, along with providing results completing those included in [12] and [13] under our test setup, we also determine the extent of the downsampling ratio that can be applied to one view before the low-resolution view starts to dominate in the perceived quality.

### 2.2 Compressed asymmetric stereoscopic video

The quantization of transform coefficients may result into perceivable coding artifacts and also often suppresses high-frequency transform coefficients and hence essentially reduces spatial resolution. Consequently, there is a tradeoff between spatial resolution of images used as input for the encoding and the quantization step size. The tradeoff between the selections of spatial resolution and the quantization step size in JPEG coding of monoscopic images was studied in [16].

Saygili et al. [17] addressed the questions what should be the level of asymmetry and whether asymmetry should be achieved by spatial resolution reduction or SNR reduction by presenting subjective assessment results. They used two test setups. The first setup included polarized glasses and a pair of projectors each having resolution of  $1,024 \times 768$ . The viewing distance was set to approximately 3 m from the screen. In the second setup, a parallax barrier auto-stereoscopic display was used. The authors concluded that when the reference view is encoded at a sufficiently high quality, the auxiliary view can be encoded above a low-quality threshold without a noticeable degradation on the perceived quality. This low-quality threshold was 31 and 33 dB in terms of average luma PSNR for the parallax barrier and the polarized projection displays, respectively. Moreover, their results showed that, at high bitrates, asymmetric coding with SNR scaling achieved the best perceived quality, while at low bitrates, asymmetric coding with spatial scaling achieved the best perceived quality. In between these two thresholds, symmetric coding was preferred over asymmetric coding.

Tam [18] compared the MR approach with a quality-asymmetric approach, in which the transform coefficients of one of the coded views were quantized coarsely. It was found that the perceived quality of the mixed-resolution videos was close to that of the higher-resolution view, while the perceived quality of the quality-asymmetric video was approximately equal to the average of the perceived qualities of the two views. The impact of the quantization of transform coefficients was verified in [15], where it was concluded that the perceived quality of coded equal-resolution stereo image pairs was approximately the average of the perceived qualities of the high-quality image and the low-quality image of the stereo pairs.

A comparison among different compression methods was presented in [19] among which MR and symmetric stereoscopic video coded with H.264/AVC were compared. Forty-seven subjects assessed 6 sequences at two bitrates typically suitable for mobile devices. The downsampling ratio of 1/2 was used for the MR bitstreams. The viewing was performed on a mobile autostereoscopic display. At the higher bitrate, symmetric stereoscopic video outperformed MR in terms of subjective acceptance and satisfaction, while the methods performed similarly at the lower bitrate.

In Sect. 4 of this paper, a systematic subjective quality evaluation test comparing different methods of asymmetric stereoscopic video coding and symmetric stereoscopic video coding are presented. The results provide some indications under which bitrates and other conditions asymmetric stereoscopic video coding is beneficial and which parameter values, such as which downsampling ratios for MR stereoscopic video, should be used. This paper therefore supplements the earlier findings reviewed above.

### 3 Extent of downsampling for mixed-resolution stereoscopic video

#### 3.1 Introduction

It is evident that there are limits on the amount of asymmetry that binocular fusion can successfully mask so that the perceived quality is closer to the quality of the higher-fidelity view. It is presumably easier to discover such limits in subjective tests when only one type of asymmetry is applied. Hence, studying uncompressed MR stereoscopic video in subjective tests makes it possible to assess such limits in resolution asymmetry between views and avoids the difficulty of analyzing the results of subjective experiences when views undergo multiple types of asymmetry. In this section, we seek to clarify as follows: “*under which viewing conditions uncompressed mixed-resolution stereoscopic video is similar to full-resolution symmetric stereoscopic video in terms of subjective quality.*” The research question was tackled by performing a subjective quality evaluation study and analyzing the results. This section extends the discussion of the subjective experiment as reported in [20] by providing more technical detail, for example, angular width, visual horizontal angle, subjective scores, and PSNR of test materials. Section 3.2 introduces the used test material, while Sect. 3.3 presents the test setup. The results are presented and analyzed in Sect. 3.4.

#### 3.2 Test material

A subjective test was performed to evaluate the subjective quality of MR stereoscopic video. The test was carried out using five sequences: undo dancer, dog, pantomime, champagne tower, and newspaper. All these sequences, presented in Fig. 2, are common test sequences in the 3D Video (3DV) ad hoc group of the moving picture expert group (MPEG). No audio track was available for any of the test sequences. The duration of all sequences in all experiments was limited to 10 s. The user perception of video quality may vary between different content types; for example, viewers may perceive action sequences differently from slow moving sequences. In order to characterize the content of the sequences, spatial and temporal perceptual information were determined using spatial information (SI) and temporal information (TI) metrics [21], although they may not always correlate well with individual’s perception experience. Considering these values, one can have a general approximation on the amount of details available in the video and how much temporal movement is expected during the content playback. The obtained SI and TI results are reported in Table 1.

For each sequence, we had the possibility to choose between several camera separations or view selections. This was studied first in a pilot test of 9 subjects. The test pro-

**Fig. 2** **a** Undo dancer, **b** dog, **c** pantomime, **d** champagne tower, **e** newspaper



**Table 1** Spatial and temporal complexity of sequences calculated using SI and TI metrics

Sequence	SI	TI
Undo dancer	98.6	23.0
Dog	90.7	23.6
Pantomime	108.3	47.0
Champagne tower	107.0	24.8
Newspaper	77.6	15.4

cedure of the pilot test was similar to that of the actual test presented in Sect. 3.3. The best average subjective viewing experience rating for undo dancer was obtained with the camera separation of 4 cm, while in the other tests, separations of 2, 6, 8, 14, and 26 cm dropped the average subjective viewing experience rating by less than 1 point on a 7-point scale. For other sequences, camera separations of 5, 10, 15, and

20 cm were tested and 5 cm separation provided the highest subjective ratings for all sequences.

Test clips were prepared as follows. Both the left and the right view image sequences were first downsampled from their original resolution to the “full” resolution presented in Table 2. The “full” resolution was selected to occupy the largest possible area on the used monitor (see Sect. 3.3) with a downsampling ratio of 1/2, 5/8, or 3/4. Moreover, the same downsampling ratio was along both directions to keep the pixel aspect ratio unchanged. To achieve the full-resolution (FR) sequences, downsampling ratio 1/2 and 3/4, were applied in both directions for undo dancer and newspaper, respectively, and 5/8 for the rest of the sequences. No cropping was applied in the conversion from the original resolution to the “full” resolution.

Two sets of test sequences were then generated, differing in whether the left view or the right view was downsampled



**Table 2** Spatial resolutions and angular widths of sequences

	Original	Full	1/2	3/8	1/4	Angular width
Undo dancer	1,920 × 1,080	960 × 540	480 × 270	360 × 202	240 × 135	40.4°
Dog	1,280 × 960	800 × 600	400 × 300	300 × 225	200 × 150	34.1°
Pantomime	1,280 × 960	800 × 600	400 × 300	300 × 225	200 × 150	34.1°
Champagne	1,280 × 960	800 × 600	400 × 300	300 × 225	200 × 150	34.1°
Newspaper	1,024 × 768	768 × 576	384 × 288	288 × 216	192 × 144	32.8°

and subsequently upsampled. In other words, in the first set of sequences, the left view was downsampled to 1/2, 3/8, or 1/4 resolution and subsequently upsampled for rendering on the display, while the right view was kept at “full” resolution. In the second set, the right view was downsampled and subsequently upsampled, while the left view was kept at “full” resolution. This arrangement of preparing two sets of sequences was done so that we could study the effect of eye dominance on the subjective quality of asymmetric stereoscopic sequences. The tested downsampling factors were 1/2, 3/8, and 1/4 symmetrically along both coordinate axes. The resolutions of the test sequences are provided in Table 2. The filters included in the JSVM reference software of the scalable video coding standard were used in the downsampling and upsampling operations [22]. The default method 0 for down and upsampling was enabled for the process. For downsampling, a sine-windowed sinc-function designed to support an extended range of spatial scaling ratios, as required by Extended Spatial Scalability (ESS), was applied. For upsampling the Scalable Video Coding (SVC), normative upsampling method designed to support ESS was applied. This filter includes a 4 tap filter with coefficients  $[-3, 19, 19, -3]$  which is originally derived from the Lanczos-3 filter. This interpolation supports any inter-layer scaling ratios, which can also be different in horizontal and vertical.

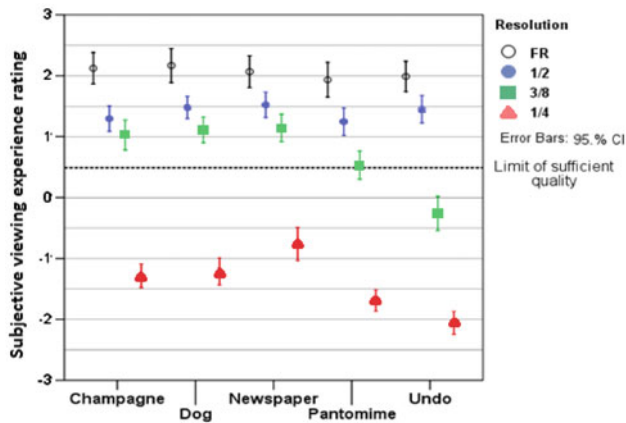
### 3.3 Test setup

The sequences were displayed un-scaled with a black background on a Hyundai P240W with a 24” polarizing stereoscopic screen having a total resolution of 1,920 × 1,200 pixels and a resolution of 1,920 × 600 per view when used in stereoscopic mode. The viewing distance was set to 70 cm because in a trial test, it yielded slightly better subjective ratings with smaller quality variation compared to those of the viewing distance of 110 cm. Since the image height was slightly different and the images were displayed un-scaled, the viewing distance of 70 cm corresponded to the range of 2.1–2.4 H for different sequences, where H is the image height. Table 3 reports the visual angle in pixels per degree (PPD) for the test setup. Moreover, Table 2 reports the angular widths in degrees.

**Table 3** Visual angle (in pixels per degree)

Downsampling ratio	Visual horizontal angle
1	22.8
1/2	11.4
3/8	7.6
1/4	5.7

Ten subjects with an average age of 21 years and without substantial prior experiences on stereoscopic video participated in the test. As we intended to confirm the previously achieved results regarding the eye dominance effect on the perceived visual quality of asymmetric stereoscopic video, half of the viewers were right-eye-dominant, while the other half were left-eye-dominant. Prior to the experiment, the viewers were subject to a thorough vision screening. The participants were screened for far and near visual acuity of each eye with a rejection criterion of 20/40 tested with Lea Numbers [23], stereoacuity criterion was 60 arcsec tested with the TNO stereo test. Criteria for near horizontal phoria, tested with the Maddox Wing test [24], were 13D for exophoria and 7D for esophoria, and 1D for vertical phoria. All participants had a stereoscopic acuity of 60 arc sec or better. The following visual tests were conducted for all participants: far and near visual acuity, stereoscopic acuity (Randot test), contrast sensitivity (Functional Acuity Contrast Test), near point of accommodation and convergence RAF gauge test [25], and the interpupillary distance. Viewers who were found not to have normal visual acuity and stereopsis were rejected. The duration of subjective test was limited to 45 min to prevent eye strain and fatigue in subjects. D50 white point, ambient illuminance level of ~200 lux, and 20% image surround reflectance were fixed as the viewing conditions of all experiments. Moreover, the background noise level was kept equal or less than 30 dBA. The subjective test started with a combination of anchoring and training. The participants were shown both extremes of the quality range of a stimulus to familiarize the participants with the test task, the contents, and the variation in quality to be expected in the actual test that followed. The test sequences were presented one at a time in a random order and appeared twice in the test session. Each sequence was rated independently after its presentation utilizing an on-screen scoring



**Fig. 3** Average subjective viewing experience ratings and the 95 % CI

scroll bar. After each rating, the next sequence started, and hence, the time used for rating was not limited in any of the experiments.

In this experiment, an integer scale in the range of  $-3$  to  $3$  was used for the rating. At the beginning of the test, the scales were presented and explained orally by the test coordinator to the participants until they understood everything thoroughly. The viewers were instructed that  $-3$  means “very bad” or “not natural,”  $0$  is “mediocre”, and  $3$  stands for “very good” or “very natural.” Moreover, the viewers were asked to estimate the limit of sufficient quality [26] with a line on the general image quality scale after viewing each test sequence. This value estimated the minimum subjective rating over which the quality was acceptable for the viewers. Observers were allowed to keep the limit of the sufficient quality at the same point for the whole experiment.

### 3.4 Results

#### 3.4.1 Limit of downsampling ratio

Figure 3 presents the average values and the 95 % confidence interval (CI) of the subjective viewing experience ratings. Furthermore, it displays the average limit of sufficient quality, which did not vary very much between sequences. It can be seen that the FR stereoscopic video sequences outperformed the MR sequences in all test cases. The quality of all MR stereoscopic image sequences downsampled by 1/2 both horizontally and vertically was clearly above the limit of sufficient quality. For three of the sequences, the downsampling ratio of 3/8 provided a quality higher than the limit of sufficient quality, while the quality of the MR sequences with the downsampling ratio of 1/4 was clearly unacceptable in terms of subjective image quality. Moreover, we observed that 70 % of the total rating interval was covered by the average subjective viewing experience ratings.

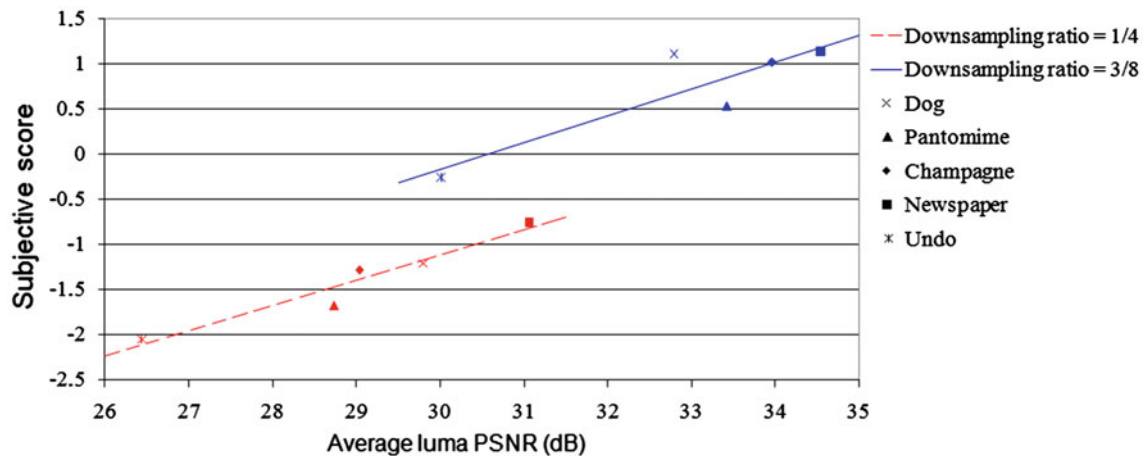
When compared to earlier studies [12, 13], the performance of the MR sequences relative to the respective FR sequences was worse. This might be explained by the chosen viewing distance in relation to the physical size of a pixel. It has also been established that when the angular resolution (e.g. in pixels per degree) stays unchanged, the greater the angular size of the display, the more contrast sensitivity the HVS has [27]. Thus, the threshold angular resolution for mixed-resolution stereoscopic video may also depend on the angular size of the display. In the viewing conditions used in this test, downsampling ratios 1/2, 3/8, and 1/4 corresponded to 11.4, 7.6, and 5.7 PPD (of viewing angle), respectively, in the lower-resolution view. As a comparison, the downsampling ratios of 1/2 and 1/4 in [12] corresponded to more than 15 and close to 10 PPD, respectively, as far as we could conclude from the information provided in the paper. The exact values for pixels per viewing angle could not be concluded from the information given in [13], but the authors discovered equivalently to our results that the subjective difference between FR and MR was a descending function of the resolution in terms of the number of pixels.

Moreover, we analyzed whether the subjective image quality ratings had any correlation to the average luma PSNR of the lower-resolution view. The downsampled views were first upsampled to the FR, and the PSNR values were derived against the FR sequences. Then, a least square estimate was derived for the relation of the subjective image quality ratings and the obtained average luma PSNR values. Finally, a Pearson’s correlation coefficient was derived between the least square estimate and the PSNR values. A large Pearson’s correlation value can be assumed to indicate that the lower-resolution view contributed more heavily to the image quality rating. Table 4 provides the PSNR of the left view and the corresponding subjective score.

A comparison between the PSNR values and the subjective viewing experience ratings of the views downsampled by ratio 1/2 resulted in Pearson’s correlation coefficient equal to 0.10, indicating that there was practically no correlation between the subjective image quality rating and the average luma PSNR of the downsampled view. The data points and the resulting least square fit for downsampling ratios 3/8 and 1/4 are presented in Fig. 4. Interestingly, the slope of the linear estimations for downsampling ratios 3/8 and 1/4 was similar and equal to 0.30 and 0.28, respectively. Along with obvious similarity of the subjective scores and the linear estimations, we further confirmed the correlation by deriving the root mean square error values, 0.25 and 0.11, and the Pearson’s correlation coefficients, 0.88 and 0.97, for downsampling ratios 3/8 and 1/4, respectively. This analysis indicates that the PSNR of the lower-resolution view correlated with subjective perception at downsampling ratios of 3/8 and 1/4. As full-reference objective quality metrics, such as PSNR, were not applicable for the full-resolution view, no analysis

**Table 4** The average luma PSNR of the left view and the average subjective viewing experience rating for different downsampling ratios

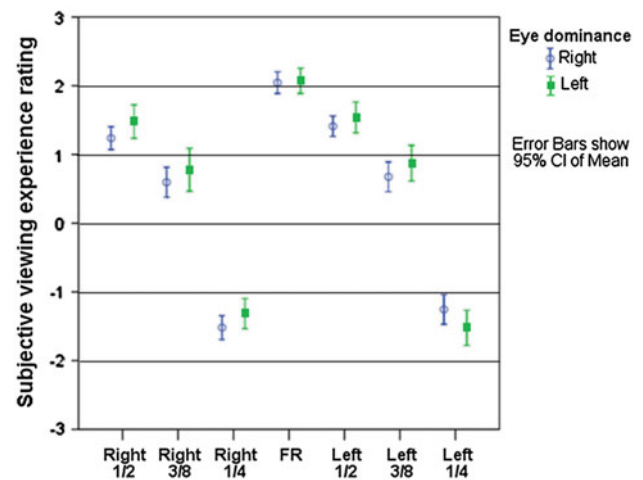
Downsampling ratio	1/2	3/8	1/4
	PSNR in dB–SSIM (average subjective rating)		
Dog	37.60–0.985 (1.47)	32.79–0.970 (1.11)	29.80–0.948 (–1.21)
Pantomime	35.62–0.990 (1.24)	33.42–0.979 (0.53)	28.74–0.965 (–1.68)
Champagne	36.32–0.993 (1.29)	33.96–0.988 (1.02)	29.04–0.983 (–1.28)
Newspaper	36.93–0.972 (1.52)	34.54–0.943 (1.14)	31.06–0.912 (–0.76)
Undo dancer	32.82–0.887 (1.45)	30.01–0.825 (–0.26)	26.44–0.778 (–2.05)

**Fig. 4** Correlation of the average luma PSNR of the lower-resolution view and the subjective viewing experience ratings, *blue*=downsampling ratio 3/8, *red*=downsampling ratio 1/4 (color figure online)

on the subjective impact of the full-resolution view was feasible with a similar method. It would therefore require further studies to verify whether the full-resolution view was dominant in the subjective quality ratings for downsampling ratio 1/2 and similarly whether the lower-resolution view was dominant at downsampling ratios 3/8 and 1/4 for the viewing conditions and the sequences used in this experiment.

### 3.4.2 Eye dominance

As explained above, there were both left- and right-eye-dominant participants in the test which included two sets of test sequences, differing in whether the left view or the right view was downsampled and subsequently upsampled. Both left and right-eye dominant subjects scored the two sets of test sequences. Figure 5 presents the average ratings given by the left- and right-eye-dominant viewers, separately. The labels of the horizontal axis identify which view was downsampled and the downsampling factor. It can be observed that there is always an overlap of the 95% confidence interval for all the respective scores, hence indicating that the eye dominance of the viewers had no significant impact on the perceived quality of the MR sequences used in the test. However, at the downsampling ratio of 1/4 along both cor-

**Fig. 5** Impact of eye dominance versus downsampled view

dinate axes, the average rating of the MR sequences where the full-resolution view was the same as the dominant eye of the viewer was slightly higher than the average rating of the other sequences of the same downsampling ratio.

We also performed statistical significance comparison achieved by the Wilcoxon signed-rank test on the results. The scores from the left- and right-eye-dominant observers

were tested against each other in order to find out whether their evaluations of the sequences differ in any case. All test cases achieved a P value equal to 1 except champagne and dog sequences at downsampling ratios of 1/2 and 3/8, respectively, for which the P values were 0.86 and 0.885, respectively. In other words, there were no significant differences of ratings between the left- and right-eye-dominant viewers based on these results. Our results therefore confirmed the earlier findings in [14] and [15] that eye dominance has no statistically significant impact on how MR sequences are rated subjectively.

## 4 Subjective quality assessment of asymmetric stereoscopic video coding

### 4.1 Introduction

Asymmetric stereoscopic video is perceived by the HVS in such a way that the lower quality of one view, due to compression artifacts, might be masked by the higher quality view. Therefore, we seek to assess the subjective quality of asymmetric stereoscopic videos with different quality combinations. For single-view video, there are a number of objective quality measures which can be used [28]. However, when it comes to stereoscopic video, objective quality assessment metrics may face some ambiguity as how to perform the joint assessment fairly, since there are two views involved with different qualities. In this section, we seek an answer to the following question: “Does asymmetric stereoscopic video coding make sense from a subjective quality point of view?” The approach to reach a conclusion is based on subjective quality assessment of symmetric and asymmetric stereoscopic videos having the same bitrate. Furthermore, the impact of downsampling ratio in mixed-resolution stereoscopic video coding is analyzed in terms of encoding computational complexity. This section further extends our preliminary results in [29].

### 4.2 Test material

The tests were carried out using four sequences: undo dancer, dog, pantomime, and newspaper. Three types of sequences were tested as follows:

1. Full-resolution with symmetric quality in both views
2. Full-resolution with asymmetric quality between the views caused by different quantization step of transform coefficients
3. Mixed-resolution with asymmetric quality

The uncompressed full-resolution sequences were generated by downsampling both the left and right view

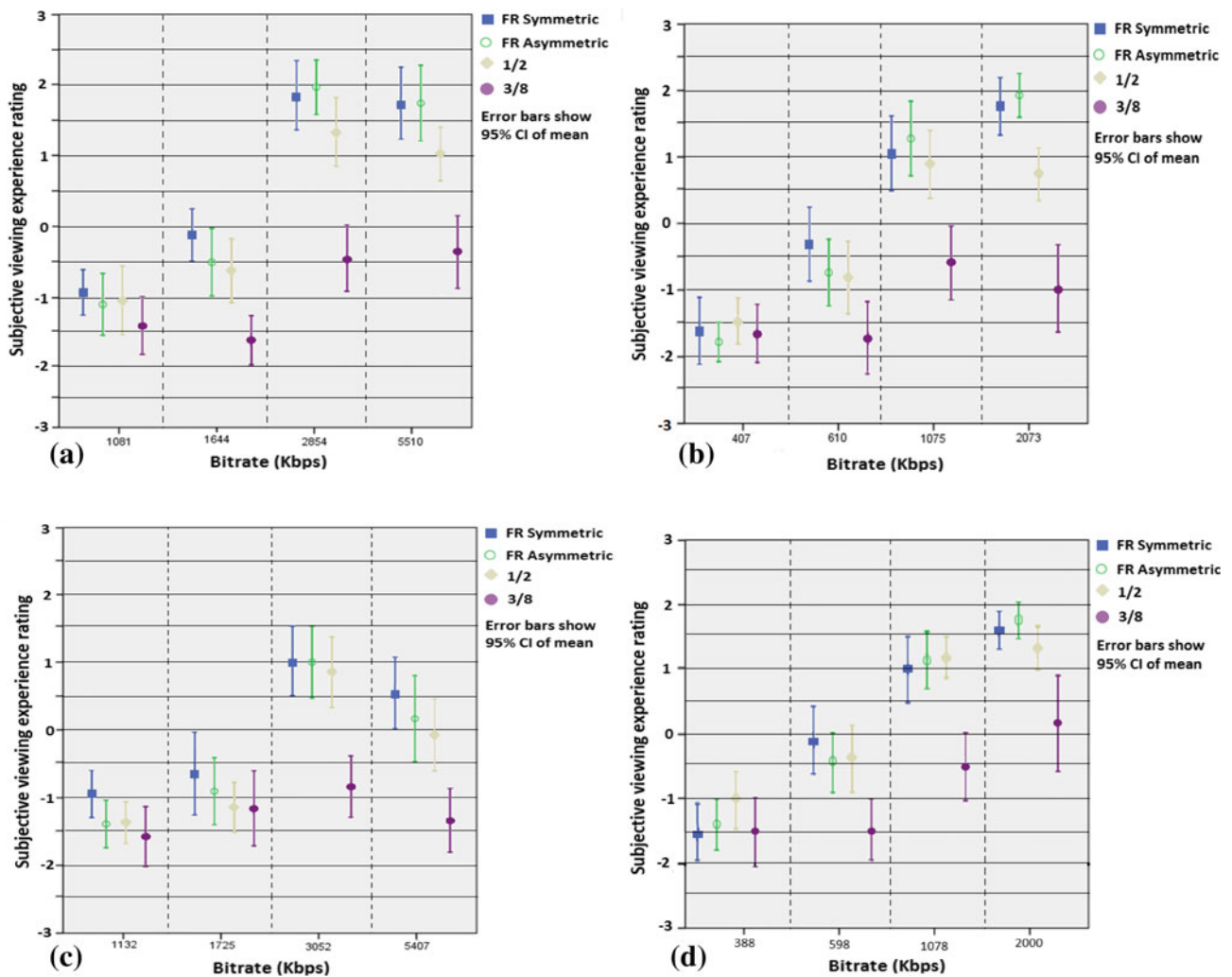
**Table 5** Spatial resolutions of different sequences

	Full	1/2	3/8
Undo dancer	960 × 576	480 × 288	360 × 216
Others	768 × 576	384 × 288	288 × 216

image sequences from their original resolution to the “Full” resolution mentioned in Table 5. The mixed-resolution uncompressed sequences were generated from the FR ones by downsampling the left view further. Downsampling ratios 1/2 and 3/8 were symmetrically applied horizontally and vertically. As in Sect. 3.4.2, we confirmed that eye dominance was not shown to have an impact which view is provided with a better quality, only one set of MR sequences was prepared. Views were independently coded using H.264/AVC in order to treat the FR and MR cases as equally as possible and prevent affecting the results by different performance of inter-view prediction depending on the downsampling ratios. Moreover, since no inter-view prediction has been standardized for a MR coding scheme, we specifically avoided the use of non-standardized codecs to provide as generally applicable results as possible. Examples of coding arrangements enabling mixed-resolution stereoscopic video with inter-view prediction have been proposed, for example, in [30] and [31].

The duration of a viewing session was limited to less than 1 h to avoid viewers becoming exhausted. Hence, the experiment was split into two sessions, where 9 and 7 naïve subjects attended the assessment tests, respectively. None of the viewers attended both sessions. Test clips having the bitrate corresponding to QP values 30 and 39 were tested in one session, whereas the remaining test clips were tested in the other test session.

The quality and bitrate of H.264/AVC bitstreams are controlled by the quantization parameter (QP). In order to get results from a large range of qualities and compressed bitrates, four constant quantization parameter (QP) values, 25, 30, 35, and 39, were selected for symmetrically compressed FR sequences. The horizontal axis of Fig. 6 displays the bitrates for different test sequences resulting from this QP value selection. A number of candidate asymmetric FR and MR bitstreams were generated, each having a bitrate within 5% of the bitrate of the corresponding symmetric full-resolution bitstream. The QP of a view was kept unchanged throughout the sequence in order to avoid any consequences of time-varying quality on the results. FR sequences with asymmetric quality were created by decreasing the QP for one view and increasing it for the other one. Table 6a presents these selected QP values. Consequently, a large variety of compressed MR combinations were considered, and the best combinations were selected in expert viewing for the actual subjective viewing test by naïve viewers. Table 6b, c summarize the QP selections for the downsampling ratio of 1/2



**Fig. 6** Results of compressed MR subjective tests for sequences: **a** undo dancer, **b** newspaper, **c** pantomime, **d** dog

and 3/8, respectively. These selections of QP values caused the bitrates of the lower-resolution view to vary from 33 to 39 % relative to the bitrate of both the views together. In addition, the uncompressed FR and MR sequences were included in the viewed sequences to obtain a reference point for the highest perceived quality of a particular sequence.

### 4.3 Results and discussion

The average subjective viewing experience ratings are presented in Fig. 6. The results of both testing sessions are merged into the same figure, even though they are not fully comparable due to different test stimuli and participants. The subjective quality of MR clips with downsampling ratio 3/8 along both axes is clearly inferior to the subjective quality of all other corresponding test cases. Thus, the results of downsampling ratio 3/8 are not discussed further. Moreover, although the confidence intervals overlap for the two highest bitrates in Fig. 6c, the average subjective ratings of the high-

est bitrate are slightly lower than the second highest bitrate. This is due to the fact that the experiment was divided to two sessions, and as a result, all four bitrates are not comparable. The highest bitrate and second lowest bitrate were included in the same session while the two other bitrates in another session.

Figure 6 indicates that mixed-resolution stereoscopic video of downsampling ratio 1/2 along both coordinate axes performed close to full-resolution symmetric stereoscopic video. Moreover, it confirms that except for the highest bitrate of newspaper, there is an overlap of the 95 % confidence intervals of the subjective ratings of FR symmetric, FR asymmetric, and MR with downsampling ratio 1/2 for each test sequence. However, the use of mixed-resolution coding can be justified in many applications by its lower computational complexity. Furthermore, it can be observed from Fig. 6 that the performance of mixed-resolution coding of downsampling ratio 1/2 depends on the input sequence to some extent.

**Table 6** QP selection (left-right) for asymmetric stereo bitstreams. a represents QP for FR asymmetric quality, while b and c represent QP selection where the left view is downsampled with ratio of 1/2 and 3/8, respectively

QP	39-39	35-35	30-30	25-25
(a) FR asymmetric bitstreams				
Undo dancer	42-36	38-32	32-28	27-23
Dog	41-37	27-33	32-28	27-23
Pantomime	42-36	37-33	33-27	28-22
Newspaper	42-36	37-33	32-28	27-23
(b) MR bitstreams with downsampling ratio of 1/2				
Undo dancer	33-36	30-32	25-28	20-23
Dog	33-37	30-33	24-28	19-23
Pantomime	34-36	31-32	24-28	20-22
Newspaper	33-36	30-32	24-28	20-23
(c) MR bitstreams with downsampling ratio of 3/8				
Undo dancer	32-36	29-32	24-28	19-23
Dog	32-36	29-32	24-27	19-22
Pantomime	32-36	29-32	24-27	19-21
Newspaper	31-36	28-32	24-27	20-22

Objective quality metrics were applied to the sequences to analyze the subjective viewing results as follows. Since to our knowledge, no widely adopted objective metrics for stereoscopic video are available, we verified the results with two common metrics: PSNR and structured similarity (SSIM) [32,33]. The average luma PSNR was derived for each view of each bitstream. For mixed-resolution bitstreams, a decoded view of a lower-resolution was upsampled before the PSNR calculation to have comparable results with full-resolution bitstreams. In the following, the PSNR of the left (L) and right (R) views of the full-resolution symmetric, full-resolution quality-asymmetric, and mixed-resolution bitstreams are marked with  $P_{SFRL}$ ,  $P_{SFRR}$ ,  $P_{AFRL}$ ,  $P_{AFRR}$ ,  $P_{MRL}$ , and  $P_{MRR}$ , respectively. SSIM values were also derived for each view of each bitstream similarly to PSNR. In the following, the SSIM values are marked in a similar fashion as, that is,  $S_{SFRL}$ ,  $S_{SFRR}$ ,  $S_{AFRL}$ ,  $S_{AFRR}$ ,  $S_{MRL}$ , and  $S_{MRR}$ .

In the case of MR stereoscopic video, both blurring and blocking are involved. We analyzed the relative contribution of the views of MR bitstreams to the overall subjective quality with both PSNR and SSIM as follows. It was assumed that the average objective quality (PSNR or SSIM) of the symmetric FR bitstreams reflects the overall subjective quality. Furthermore, we assumed that when a weighted average of the objective quality values between the left and right view of an MR bitstream matches the average objective quality of the respective symmetric FR bitstream having the same subjective quality rating, the weights for the weighted averaging reveal the relative contribution of left and right views to the subjective quality. In other words, for those MR bitstreams

that had approximately equal subjective quality as the respective FR bitstreams, we derived weights  $W$  that minimized the mean square error of the difference between the weighted average of the objective quality of the left and right views and that of the FR:

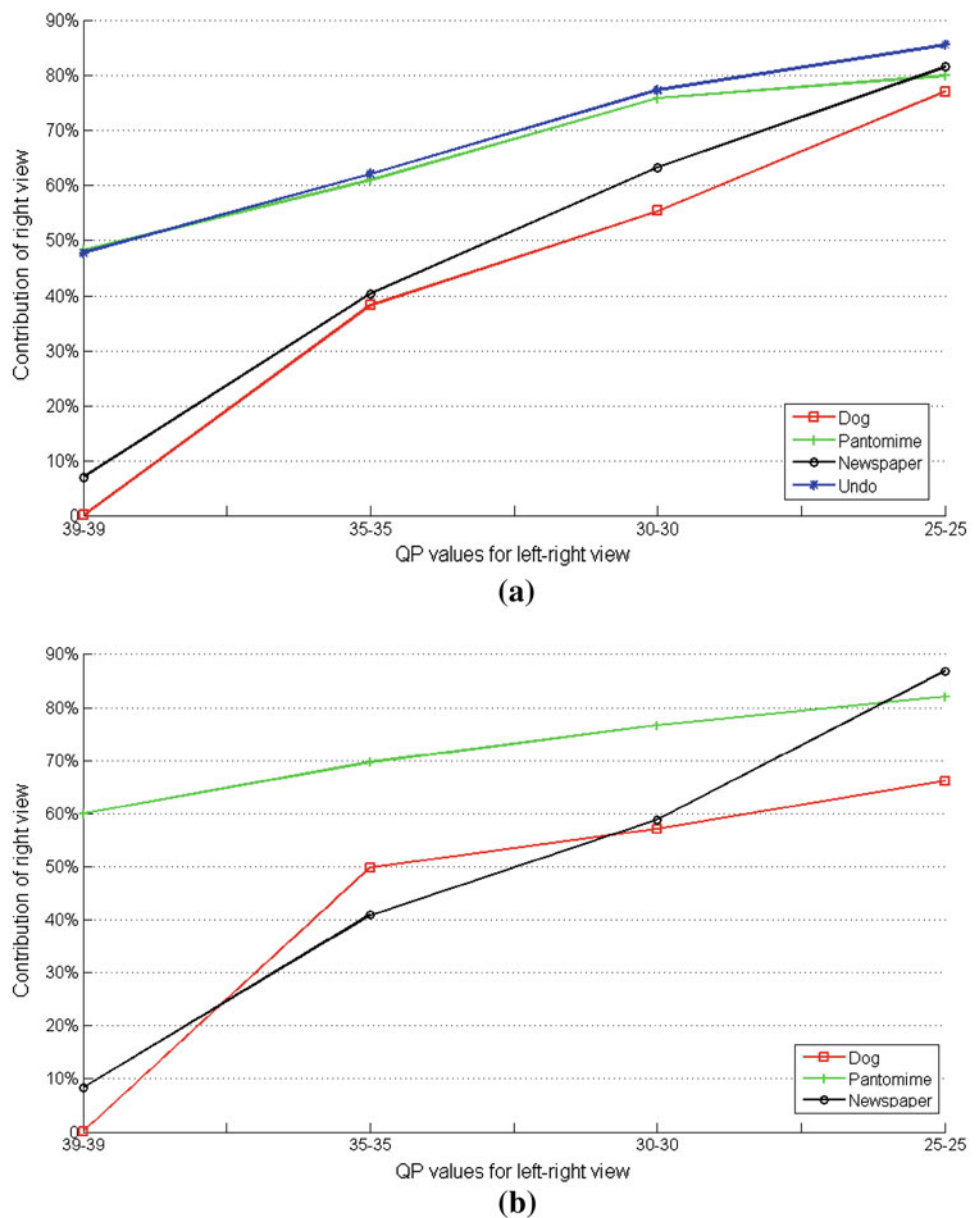
$$\text{mse} = (W \times P_{MRR} + (1 - W) \times P_{MRL} - P_{SFR})^2 \quad (1)$$

In Eq. (1),  $W \times P_{MRR} + (1 - W) \times P_{MRL}$  reflects the weighted average of MR bitstreams and mse is minimized by changing the weight ( $W$ ) over the quality of left and right views. Assuming that  $P_{MRL} < P_{SFR} < P_{MRR}$ , which is typically true because only the left view is downsampled and due to the downsampling, the right view gets a lower QP value compared to the right view of symmetric FR, the above expression reaches its minimum when

$$W = (P_{SFR} - P_{MRL}) / (P_{MRR} - P_{MRL}) \quad (2)$$

The same reasoning can be applied for SSIM. Figure 7a, b indicate the contribution of the right view to the overall quality, that is,  $W$ , for different QP values and sequences, derived from PSNR and SSIM, respectively. The results of undo dancer were not included in Fig. 7b because the MATLAB implementation of the SSIM index, utilizing the suggested empirical formula [33], seemed to fail in estimating its subjective quality. SSIM provided very close values for the left and right views for undo dancer as derived from Eq. (2). A full 100% contribution was assigned to the right view for the three highest QP values. This was not the case for the

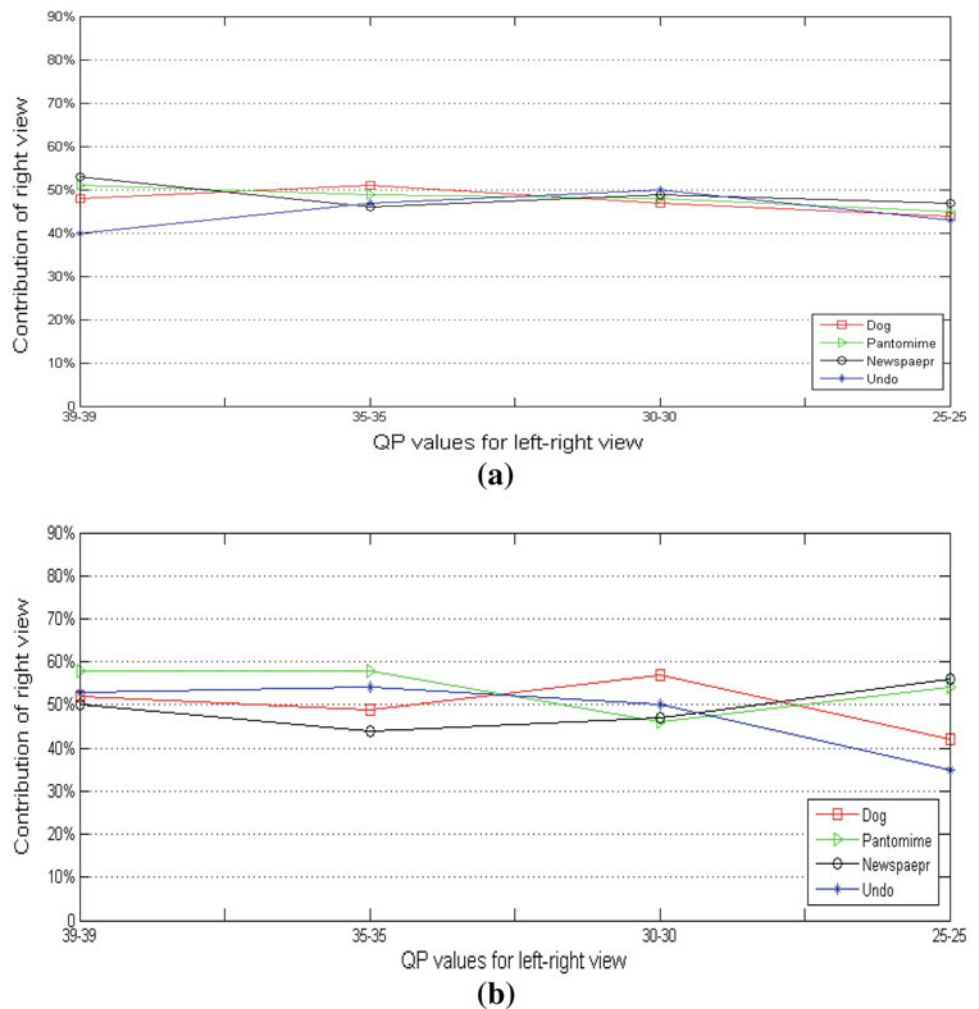
**Fig. 7** Contribution of the FR view (*right*) to the overall quality of mixed-resolution stereoscopic video measured by **a** PSNR **b** SSIM, that is, the value of  $W$  as derived with Eq. (2)



other sequences, perhaps due to the synthetic nature of the undo dancer sequence. It can be seen from Fig. 7a, b that the contribution of the right view increased when blocking decreased and that the higher the QP value became, the more contribution the left view had on the overall quality. Moreover, Fig. 7 appears to be in agreement with the conclusions in [7] that the perceived quality of the mixed-resolution videos was close to that of the higher-resolution view. This behavior was not biased by QP selection for the left and the right view for different bitrates since as reported in Table 6b, the QP difference between the left and the right view for all MR videos was kept equal or close to three. It can also be seen in Fig. 7 that the relative contribution of the right view was dependent on the sequence.

The average luma PSNR over both views of the quality-asymmetric full-resolution bitstreams, that is,  $(P_{AFRL} + P_{AFRR})/2$ , was found to be very close to that of the symmetric full-resolution bitstreams, that is,  $P_{SFR} = (P_{SFRl} + P_{SFRr})/2$ , the absolute difference being only 0.1 dB on average. The same analysis for SSIM metric resulted in an absolute difference of 0.005 on average. This finding is aligned with the earlier conclusions in [7] and [15] that the perceived quality of the quality-asymmetric video was approximately the mean of the perceived qualities of the two views. The same analysis, as reported for MR stereoscopic video in Fig. 7, was performed for quality-asymmetric full-resolution sequences. The results are provided in Fig. 8 for both PSNR and SSIM objective metrics showing that both

**Fig. 8** Contribution of the right view to the overall quality of quality-asymmetric full-resolution stereoscopic video measured by **a** PSNR **b** SSIM



views contributed almost equally to the final quality of the stereoscopic video.

As discussed above, MR coding did not provide a better subjective quality compared to FR coding. However, due to the smaller spatial resolution, the use of MR coding may be justified. A complexity comparison for encoding the full and lower-resolution views in our experiments is presented in Fig. 9. The experiments were performed on Windows OS with a dual-core CPU having a clock rate of 3.16 GHz. The execution time for the FR view consisted of the encoding time, and for the lower-resolution view, it included both the encoding and the downsampling times. Since the encoding time varied depending on the ongoing processes of the PC, an average value of seconds per frame over five different QP values for full-length videos was calculated. As illustrated in Fig. 9 by decreasing the spatial resolution by ratio 1/2 and 3/8 both vertically and horizontally, the encoding time decreased on average to 36 and 21% of the encoding time for the FR sequences, respectively.

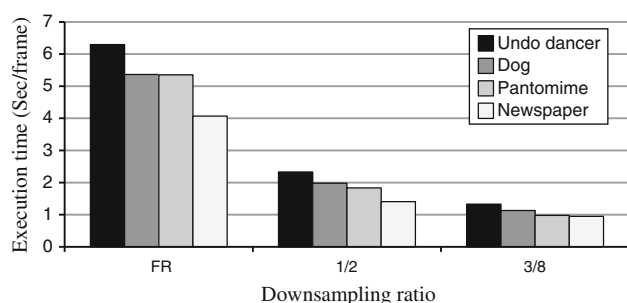
To reduce the amount of time-taking subjective experiments, it is preferred to estimate the subjective quality of

asymmetric stereoscopic video by a reliable model depending on available information, for example, the characteristics of the viewing conditions, the used asymmetric coding scheme, and the viewed video content. In [34], we tried to estimate the subjective quality of asymmetric stereoscopic video taking into account the number of pixels per degree of viewing angle. The results showed high correlation between subjective ratings and pixels per degree values but were obtained with a relatively small amount of subjective test data. In order to verify the results of [34] and to develop the model further, we plan to conduct extensive subjective tests under multiple test setup conditions, different asymmetric coding schemes, and various video clips.

## 5 Conclusions

In this paper, we attempted to discover suitable methods and configurations for asymmetric stereoscopic video coding through two sets of systematic subjective quality evaluation experiments. We studied the subjective impact of downsam-





**Fig. 9** Encoding time comparison for FR view and downsampled views

pling applied for one of the views in an uncompressed mixed-resolution (MR) stereoscopic video. In our experiment, FR sequences always outperformed MR sequences. However, the quality of the MR sequences where one view was downsampled by a factor of 1/2 horizontally and vertically was clearly acceptable. We found that the lower-resolution view appeared to become dominant in the subjective quality rating at a certain downsampling ratio, which seemed to depend on the sequence, the angular resolution, and the angular width.

A subjective test comparing symmetric full-resolution, quality-asymmetric full-resolution, and mixed-resolution stereoscopic video coding was also presented. The performance of symmetric and quality-asymmetric full-resolution bitstreams was found to be approximately equal. Mixed-resolution stereoscopic video with downsampling ratio 1/2 along both coordinate axes performed similarly to the full-resolution bitstreams in most of the test cases. Due to the lower required processing complexity, the use of mixed-resolution stereoscopic video can be considered in many applications. Mixed-resolution stereoscopic video with downsampling ratio 3/8 along both coordinate axes was found to be clearly inferior to all other tested coding arrangements and did not yield acceptable quality at any bitrate.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

## References

- Chen, Y., Wang, Y.-K., Ugur, K., Hannuksela, M.M., Lainema, J., Gabbouj, M.: The emerging MVC standard for 3D video services. *EURASIP J. Adv. Signal Process.* **2009**, 13 (2009); Article ID 786015. doi: 10.1155/2009/786015
- ITU-T Recommendation H.264.: Advanced Video Coding for Generic Audiovisual Services (March 2009)
- Merkle, P., Smolic, A., Muller, K., Wiegand, T.: Efficient prediction structure for multiview video coding. *IEEE Trans. Circuits Syst. Video Technol.* **17**(11), 1461–1473 (2007)
- Sullivan, G.J., Ohm, J.-R., Han, W.-J., Wiegand, T.: Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits Syst. Video Technol.* **22**, 1649–1668 (2012)

- Suzuki, T., Hannuksela, M.M., Chen, Y., Hattori, S., Sullivan, G.J. (eds.): MVC extension for inclusion of depth maps draft text 4. Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1001 (July 2012)
- Suzuki, T., Hannuksela, M.M., Chen, Y., Hattori, S., Sullivan, G.J. (eds.): MVC extension for inclusion of depth maps draft text 6. Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-C1001 (Mar. 2013)
- Blake, R.: Threshold conditions for binocular rivalry. *J. Exp. Psychol. Human Percept. Perform.* **3**(2), 251–257 (2001)
- Julesz, B.: *Foundations of Cyclopean Perception*. University of Chicago Press, Chicago (1971)
- Perkins, M.G.: Data compression of stereopairs. *IEEE Trans. Commun.* **40**(4), 684–696 (1992)
- Aksay, A., Bilen, C., Bozdagi Akar, G.: Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images. In: 13th European Signal Processing Conference, EUSIPCO-2005, Turkey (Sep. 2005)
- Aflaki, P., Hannuksela, M.M., Hakala, J., Häkkinen, J., Gabbouj, M.: Joint adaptation of spatial resolution and sample value quantization for asymmetric stereoscopic video compression: a subjective study. In: Proceedings of the International Symposium on Image and Signal Processing and Analysis (Sep. 2011)
- Stelmach, L., Tam, W.J., Meegan, D., Vincent, A.: Stereo image quality: effects of mixed spatio-temporal resolution. *IEEE Trans. Circuits Syst. Video Technol.* **10**(2), 188–193 (2000)
- Brust, H., Smolic, A., Mueller, K., Tech, G., Wiegand, T.: Mixed resolution coding of stereoscopic video for mobile devices. In: Proceedings of the of 3DTV Conference (May 2009)
- Meegan, D.V., Stelmach, L.B., Tam, W.J.: Unequal weighting of monocular inputs in binocular combination: implications for the compression of stereoscopic imagery. *J. Exp. Psychol. Appl.* **7**(2), 143–153 (2001)
- Seuntiens, P., Meesters, L., Ijsselstein, A.: Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Trans. Appl. Percept.* **3**(2), 96–109 (2006)
- Bruckstein, A.M., Elad, M., Kimmel, R.: Down-scaling for better transform compression. *IEEE Trans. Image Process.* **12**(9), 1132–1144 (Sep. 2003)
- Saygili, G., Gürler, C.G., Tekalp, A.M.: Quality assessment of asymmetric stereo video coding. In: Proceedings of the of IEEE International Conference on Image Processing (Sep. 2010)
- Tam, W.J.: Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV. In: Joint Video Team document JVT-W094 (Apr. 2007)
- Strohmeier, D., Tech, G.: Sharp, bright, three-dimensional: open profiling of quality for mobile 3DTV coding methods. In: Proceedings of the SPIE International Society for Optical Engineering, vol. 7542 (Jan. 2010)
- Aflaki, P., Hannuksela, M. M., Häkkinen, J., Lindroos, P., Gabbouj, M.: Impact of downsampling ratio in mixed-resolution stereoscopic video. In: Proceedings of the of 3DTV Conference (June 2010)
- ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications (1999)
- JSVM Software: [http://ip.hhi.de/imagecom\\_G1/save/download/SVC-Reference-Software.htm](http://ip.hhi.de/imagecom_G1/save/download/SVC-Reference-Software.htm)
- Hyvärinen, L.: Lea Numbers 15-Line distance chart instructions. Web page: [leatest.net/en/vistests/instruct/2711/index.html](http://leatest.net/en/vistests/instruct/2711/index.html). Accessed 01 Jan 2011, Created 2009
- Rosenfield, M., Logan, N. (eds.): *Optometry: Science, Techniques and Clinical Management*. Elsevier, Amsterdam (2009)
- Neely, J.C.: The R.A.F. near-point rule. *British J. Ophthalmol.* **40**(10), 636–637 (Oct. 1956)

26. Nyman, G., Häkkinen, J., Koivisto, E.-M., Leisti, T., Lindroos, P., Orenius, O., Virtanen, T., Vuori, T.: Evaluation of the visual performance of image processing pipes: information value of subjective image attribute. In: Proceedings of the SPIE, vol. 7529 (Jan. 2010)
27. Barten, P.G.J.: The effects of picture size and definition on perceived image quality. *IEEE Trans. Electron. Devices* **36**(9), 1865–1869 (Sep. 1989)
28. You, J., Reiter, U., Hannuksela, M.M., Gabbouj, M., Perkis, A.: Perceptual-based quality assessment for audio-visual services: a survey. *Signal Process. Image Commun.* **25**(7), 482–501 (2010)
29. Aflaki, P., Hannuksela, M.M., Häkkinen, J., Lindroos, P., Gabbouj, M.: Subjective study on compressed asymmetric stereoscopic video. In: Proceedings of IEEE International Conference on Image Processing (ICIP) (Sep. 2010)
30. Chen, Y., Liu, S., Wang, Y.-K., Hannuksela, M.M., Li, H., Gabbouj, M.: Low-complexity asymmetric multiview video coding. In: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME) (June 2008)
31. Brust, H., Tech, G., Mueller, K., Wiegand, T.: Mixed resolution coding with interview prediction for mobile 3DTV. In: Proceedings of 3DTV Conference (June 2010)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
33. <https://ece.uwaterloo.ca/~z70wang/research/ssim/>
34. Aflaki, P., Hannuksela, M.M., Hakala, J., Häkkinen, J., Gabbouj, M.: Estimation of subjective quality for mixed-resolution stereoscopic video. In: Proceedings of the of 3DTV-Conference (May 2011)

[P4] **P. Aflaki**, D. Rusanovskyy, M. M. Hannuksela, and M. Gabbouj; “Unpaired multiview video plus depth compression,” IEEE Digital Signal Processing, Santorini, Greece, July, 2013.

© IEEE, 2013, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# Unpaired Multiview Video Plus Depth Compression

Payman Aflaki Moncef Gabbouj  
Department of Signal Processing  
Tampere University of Technology  
Tampere, Finland  
(payman.aflaki, moncef.gabbouj)@tut.fi

Dmytro Rusanovskyy, Miska M. Hannuksela  
Nokia Research Center  
Nokia Corp.,  
Tampere, Finland  
(dmytro.rusanovskyy, miska.hannuksela)@nokia.com

**Abstract**— Recent developments of three-dimensional (3D) video coding greatly rely on the use of Multiview Video plus Depth (MVD) data format for representing a 3D scene. This type of data can be coded with conventional video compression schemes and can enable advanced 3D video functionality, such as rendering of virtual views at the decoder side. The MVD represents a 3D scene from different viewing angles as video and pixel-wise associated depth data. In this paper we consider the redundancy of depth data in the MVD representation and propose a novel scheme, called Unpaired MVD (UP-MVD) format to be used in 3D video applications. Being a subset of the MVD this new format assumes that a reduced number of depth views compared to the number of texture views can reduce bitrate as well as the encoding/decoding complexity while still providing the 3DV functionality in many scenarios. The simulation results show that the proposed unpaired MVD format outperforms the MVD format on average from 0.9% to 6.95% of Bjontegaard delta bitrate (dBR) for the baseline disparity adjustments from 50% to 100% of the coded baseline, respectively. Moreover, UP-MVD provides equal or better rate-distortion results for all test sequences for up to 20% view separation adjustment, and in five out of seven sequences a better rate-distortion performance is observed when 50% view separation adjustment is applied.

**Keywords**-3DV; MVD; View synthesis.

## I. INTRODUCTION

The presence of depth data at the decoder side enables a more flexible display of three-dimensional (3D) video compared to conventional stereoscopic and multiview video coding. While coding of two texture views provides a basic 3D perception on stereoscopic displays, it has been discovered that disparity adjustment between views is needed for adapting the content on different displays and viewing conditions as well as for individual preferences [1]. Moreover, in autostereoscopic displays a large number of views is typically required to be displayed simultaneously. However, it is impossible or impractical to transmit a large number of views through today's networks, such as the Internet, using existing video compression standards, such as the Multiview Video Coding (MVC) extension of the Advanced Video Coding (H.264/AVC) standard [2]. Therefore, the required views have to be generated in the playback device from the received views. These needs can be addressed by representing a 3D scene with a multiview video plus depth (MVD) format [3] and using the decoded MVD data

as source for depth image-based rendering (DIBR) [4]. In the MVD format each video data pixel is associated with a corresponding depth map value from which new views can be synthesized using any DIBR algorithm in the post processing stage before displaying the content.

The Moving Picture Experts Group (MPEG) issued a Call for Proposals (CfP) on 3D video coding technology in March 2011 [5]. The CfP aimed at starting the standardization of a coding format that supports advanced stereoscopic display processing and improved support for auto-stereoscopic multiview displays. As a result of the CfP evaluation [6], the MPEG and, since July 2012, the Joint Collaborative Team on 3D Video Coding (JCT-3V) [7] have initiated the development of an MVC extension to include depth maps [2], abbreviated as MVC+D, and to specify the encapsulation of coded MVD data into a single bitstream [8]. According to this specification, MVC coding [2] is applied independently to both texture and depth components of MVD, and the texture views of MVC+D bitstreams can be decoded with a conventional MVC decoder. A reference test model of MVC+D is implemented in 3DV-ATM reference software [9] and it was used in our work.

DIBR enables the projection of a texture view to a virtual viewing position. However, DIBR has intrinsic limitations of being unable to render samples in areas that become uncovered in synthesized views (termed “dis-occlusions” or “holes”) in the capturing process. This mainly happens while extrapolating one view, i.e., while rendering a view in a specific location from information of a single view. To overcome this, view interpolation can be performed in DIBR algorithms, i.e., a middle view can be rendered by exploiting information of the left- and right-side views while performing projection from different directions to fill the dis-occluded parts. The dis-occlusion problem of DIBR is due to areas covered by objects in the reference view which appear in the synthesized view. Such holes and dis-occluded areas have neither a certain depth or texture attribute, nor a correspondence in the reference views. In DIBR, these holes need to be filled properly otherwise annoying artifacts will appear in the dis-occluded regions. To solve the problem of dis-occlusions in DIBR, several algorithms have been developed. There has been an extensive research proposing different hole-filling methods [10] using simple and sophisticated image processing techniques. Most of these conventional methods

exploit neighbor pixel values to fill-in the holes by extrapolation, linear interpolation, or diffusion techniques.

In our work, we assumed that the conventional MVD representation with pixel-wise correspondence between texture and depth can be redundant for some 3D scenes and for some of use cases. It is assumed and proved in this paper that in many application scenarios, the amount of depth map data describing the 3D scene can be reduced without a significant impact on DIBR performance. Analogously to [11], [12] and [13], where it was shown that depth information can be spatially decimated within a single view improving the performance of DIBR, we consider that depth data for some views can be completely ignored since it does not present a considerable amount of information in addition to other presented views.

As a result of our study, we propose a novel Unpaired MVD (UP-MVD) format to be used in 3D video applications. Being a subset of the MVD data format, UP-MVD reduces the number of depth views compared to texture views while maintaining a sufficient view synthesis capability for typical 3DV applications and reducing the encoding/decoding complexity. The process of removing the redundant depth components within MVD data can be conducted at the post-production stage of 3D video content capturing or at the encoder side of the 3DV coding chain.

The rest of paper is organized as follows. Section II describes the proposed unpaired MVD scheme. The test material and simulation results are presented in Section III, while Section IV concludes the paper.

## II. PROPOSED UNPAIRED MVD SCHEME

### A. Motivation

MVD represents a 3D scene from different viewing angles as video and pixel-wise associated depth data. However, as demonstrated in many of the responses to the 3DV CfP [5], coding of depth map data at a reduced resolution is a viable solution for improving the rate-distortion performance of the complete 3D video coding system. As a result, for most of available MVD content, e.g. MPEG 3DV Test Set [14], the depth component can be spatially decimated within each view without a significant impact on the performance of DIBR-based 3D Video applications. As a follow up of this concept, we considered view-level decimation of the depth data, and studied the redundancy within the currently available MVD content confirming that the number of depth data compared to texture views can be reduced without sacrificing the quality of synthesized views. As a result of the analysis of practical scenarios of 3D scene capturing and 3D video applications, a few examples where a complete MVD representation is unnecessary and/or redundant are presented below.

A typical 3D video scene capturing process nowadays features a stereo camera and depth information is derived through disparity estimation process having a stereoscopic video as input. A disparity search is performed in a one-directional manner, i.e. for each sample in the first view component (e.g. a picture of the

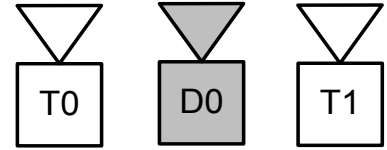


Figure 1. Stereo camera (T0 and T1) with a single ranging device (D0)

left view), the corresponding sample in the second view component (e.g. a picture of the right view) is searched. The resulting disparity picture can be converted to a depth component of the first view. The depth information for the second view can be produced by inverting the disparity vectors computed for the first view. Thus, there is no need to produce and code depth component for the second view, since this information would be completely redundant.

Alternatively, depth information may be generated using a specific depth sensor [15] [16] rather than generated in a per-pixel depth estimation process based on texture views. In such a camera setup, it is typical that a depth sensor is not collocated with any of the utilized image sensors. A visualization of this concept is shown in Fig. 1 where two cameras are accompanied by a single ranging device which is not located in the same place as any of the image sensors.

In some practical 3D Video applications, a stereo baseline adjustment would be required in relatively close proximity to one decoded texture view, whereas another decoded texture view would be displayed as it is. Therefore, a depth data associated with the first view can be sufficient as the input for DIBR in extrapolation mode and no need to process depth data associated with the second view. This example is illustrated in Fig. 2 where four stereo baselines can be achieved depending on the use of L, L1, L2 or L3 as the left view. Hence, even if a depth view is available for encoding for each texture view, this complete MVD representation for 3D scene may be unnecessary for enabling many 3D video applications and the encoder side may be

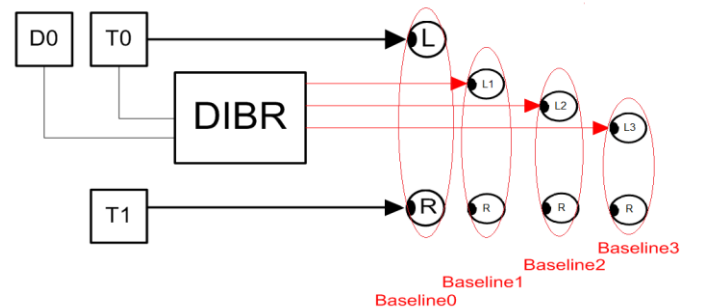


Figure 2. Visualization of depth perception adjustment by view synthesis in close proximity from one decoded view T0 and T1 are input texture views while D0 is the only input depth view L1, L2, and L3 present the synthesized views

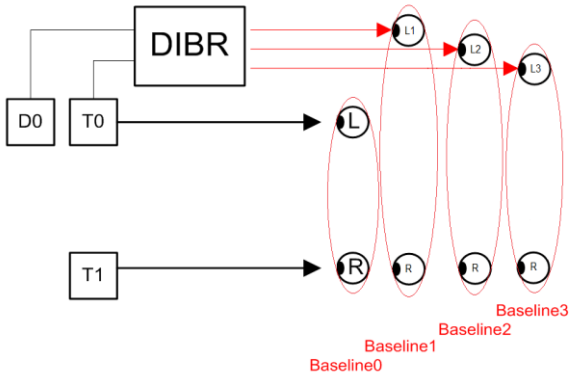


Figure 3. Visualization of depth perception adjustment by view synthesis to an extrapolated view  
T0 and T1 are input texture views while D0 is the only input depth view  
L1, L2, and L3 present the synthesized views

adjusted to limit the number of depth views to be encoded.

The need for synthesizing only one view from stereoscopic video also applies for use cases when the extrapolation of virtual views rather than interpolation is required. For example, an optimal disparity for a two-view autostereoscopic display for handheld use is typically wider than that for a living-room polarized or a shutter-glass display. Depth-enhanced stereoscopic content for handheld devices could therefore have two texture views and only one depth view, as the extrapolation quality would remain the same as that for two texture and depth views. This is depicted in Fig. 3.

As stated in Section I, DIBR does not guarantee a full and correct projection of one view to the target location, because some parts of the projected view may be dis-occluded or contain holes. The percentage of hole pixels in a rendered view depends on many factors e.g. scene characteristics and the view separation between the original and projected views. In the following paragraph, we evaluate the forward projection process on the first frame of depth views of the 2-view sequences used in this paper in order to demonstrate that the two depth views are correlated

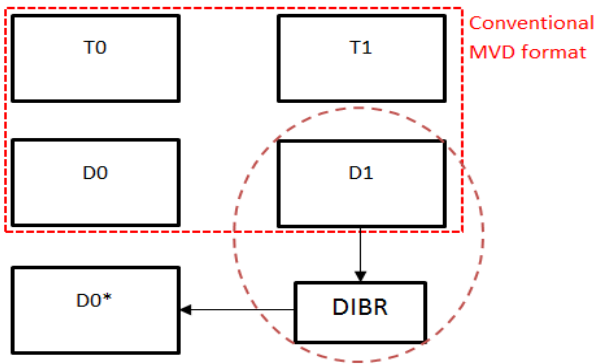


Figure 4. Forward projection of depth maps to calculate percentage of hole pixels in the rendered view

and that the second depth view contains only a moderate proportion of pixels that could not be projected from the first depth view.

For simplicity, the case with two texture views and their associated depth maps is considered (2-view MVD data). However, this approach can be extended to be applied for MVD content with more than two views. Considering Fig. 4, D1 is projected forward to the position of view 0 to produce D0\*. The missing information in D0\* results either from a different field of view compared to D0 or from the inaccuracy of the depth estimation and the depth projection algorithms. The location of the missing information is marked as a “hole”. Table I shows the proportions of pixel locations with such missing information normalized by the number of pixels in a depth image D0 for the first frame of each sequence. The estimates have been evaluated with the MPEG 3DV test set [14]. As reported in Table I, on average more than 95% of the pixel locations in D0\* contain projected pixels from D1 and hence, D0\* can be assumed to provide a proper estimation of D0. This is further studied in Section III with a series of simulations and reported objective results.

Moreover, the encoder and the decoder complexity are reduced as a direct result of the view reduction introduced in proposed MVD format compared to the conventional MVD format.

### B. Proposed schemes

In this sub-section, we describe a novel 3D video processing scheme that is based on the unpaired MVD format. The input of the proposed scheme is a conventional MVD format with a reduced number of coded depth views.

After removing some of the depth views, the full set of texture views and a subset of the depth views is coded with a modified MVC+D codec [9], resulting in a smaller bitrate than that of the respective complete MVD data coded with MVC+D. At the decoder side, the UP-MVD data is used for rendering virtual views with DIBR and the quality of these views is used for the performance evaluation. In our work we studied the impact of the use of unpaired MVD with two schemes, the details of which are given below.

Scheme 1 targets low-complexity encoder and decoder operations with reduced memory requirements, whereas Scheme

TABLE I. PERCENTAGE OF DISOCCLUDED AND UN-PROJECTED PIXELS IN D0\*

Sequence	
Poznan Hall2[17]	3.8%
Poznan Street	2.2%
Undo Dancer	3.4%
Ghost Town Fly	4.3%
Kendo	11.4%
Balloons	5.4%
Newspaper	3.5%
<b>Average</b>	<b>4.9%</b>

2 assumes a more advanced decoder providing a higher subjective and objective quality.

**Scheme 1:**

A flowchart of this scheme is depicted in Fig. 5. First, two texture views and one depth view are encoded and decoded. Then, rendered views are created with an extrapolation of the base view (T0 and D0). The stereoscopic image-pair is made of one coded texture view (T1) and for the other view based on the desired baseline, either the other coded view (T0) or one of the synthesized views (L1, L2, L3) is used. This approach has a low computational complexity and memory requirement due to the fast extrapolation from one view. However, the drawback of this scheme is the omission of T1 in the rendering process where it can significantly improve the quality of the synthesized views.

**Scheme 2:**

This scheme is similar to Scheme 1 but for the rendering process, both texture views are utilized. The corresponding flowchart of Scheme 2 is shown in Fig. 6. In this scheme, D1\* is produced from the available D0 through a projection to view location 1 followed by basic hole filling [10]. The desired views (L1, L2, and L3) are then obtained by interpolation using T0 and D0 as well as T1 and D1\*. This enables exploiting the texture information of T1 associated with D1\* in the rendering process,

and therefore, improves the quality of view synthesis compared to that achievable with Scheme 1.

III. SIMULATION RESULTS

The simulation results run under the specifications of C2 scenario of 3DV Common Test Conditions (CTC) [14] and the complete set of MPEG 3DV test sequences was utilized. In this scenario two depth-enhanced texture views are encoded and then several possible intermediate views are synthesized in-between to be exploited in stereoscopic image-pair creation. 3DV-ATM software configured in MVC+D was utilized for coding UP-MVD, and 3DV VSRS [18] was used for the rendering of virtual the depth and texture views.

In our experiments, stereo baseline adjustment was enabled in the proximity of one of the decoded texture view, whereas the other decoded texture view was used as the second view of the displayed stereoscopic image-pair. Rendered views were located at 0.1, 0.2, and 0.5 of the baseline named L1, L2, and L3, respectively, as it is shown in Fig. 5. The specific views used at the input and output of this experiment are listed in Table II.

The quality of the stereoscopic video was compared to the anchor case where both views were coded with their associated depth map. To evaluate the performance objectively, the Peak

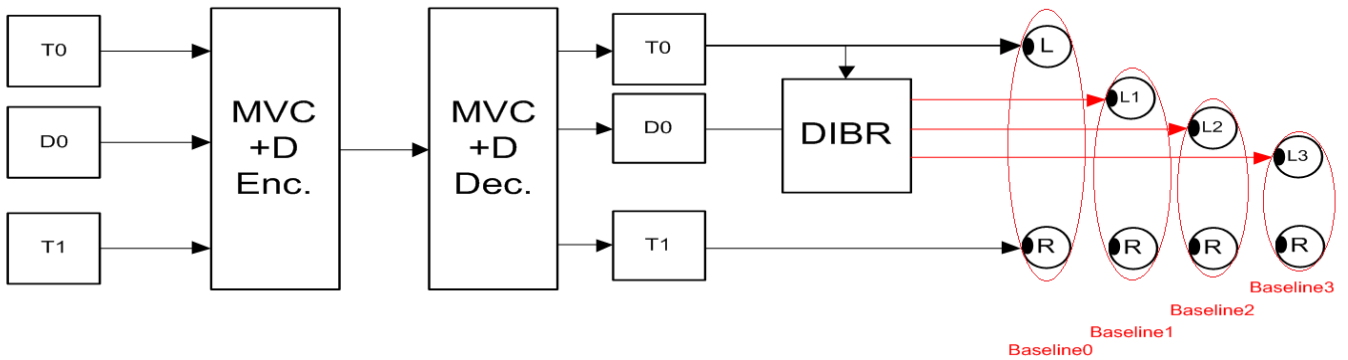


Figure 5. Proposed Scheme 1 for Unpaired MVD

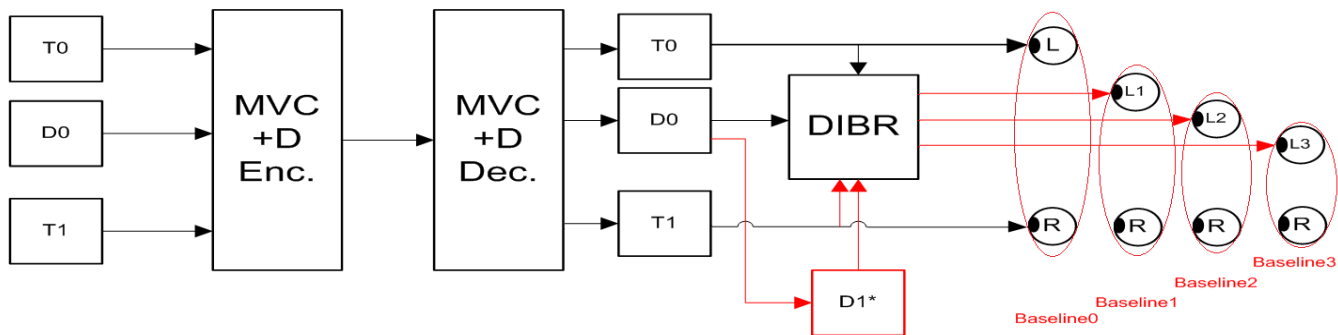


Figure 6. Proposed Scheme 2 for Unpaired MVD

TABLE II. 3DV TEST SEQUENCES, INPUT AND SYNTHESIZED VIEWS

Sequence	Input views	Synthesized views
PoznanHall2	7-6	7.5 – 7.8 – 7.9
Poznan Street	5-4	5.5 – 5.8 – 5.9
Undo Dancer	1-5	3 – 4.2 – 4.6
Ghost Town fly	9-5	7 – 5.8 – 5.4
Kendo	1-3	2 – 3.6 – 3.8
Balloons	1-3	2 – 3.6 – 3.8
Newspaper	2-4	3 – 4.6 – 4.8

Signal-to-Noise Ratio (PSNR) of both views in the stereoscopic image pairs (including one coded view and one synthesized view) was calculated. The PSNR of the synthesized views was calculated against the reference synthesized views created from the conventional MVD format including the original uncompressed texture and depth views as specified in MPEG CTC [14]. The objective results are presented using Bjontegaard delta bitrate (dBR) and delta PSNR [19]. The delta bitrate (dBR) is presented for the coded views and the stereoscopic image pairs created with one decoded view and one synthesized view. The results for the first and the second proposed schemes are provided in Table III and IV, respectively. Moreover rate-distortion curves for the largest baseline stereoscopic image pair (base + 0.9) for Scheme 2 is depicted in Fig. 7. The results show that a bitrate reduction, compared to the anchor, of 6.95% for coded views is achieved due to the removal of one depth view. Moreover, in Scheme 1, a larger baseline (i.e. base + 0.9) provides 4% dBR gain with a visible degradation in the performance when decreasing the baseline. However, this problem is addressed in Scheme 2 where a higher average performance compared to the

anchor is always achieved. As reported in Table IV, the smallest baseline achieves 0.9% of dBR compared to the anchor. This gain increases up to 5.44% for the largest baseline achieved from one decoded and one synthesized view. It can also be concluded that using Scheme 2, up to 20% view separation adjustment can be achieved while the proposed UP-MVD format always outperforms the anchor conventional MVD format. Moreover, in five out of seven test sequences in the smallest baseline configuration, the anchor is outperformed by the proposed method.

#### IV. CONCLUSIONS

In this paper we studied the objective quality of encoded and synthesized views from an MVD data format. Our assumption was that the number of depth views can be smaller than the number of texture views. Hence, the proposed UP-MVD format as a subset of MVD data format was introduced, where the number of texture views differs from the number of depth views, e.g. two texture views are accompanied with only one depth view. The proposed data format succeeded to outperform the conventional MVD data format on average by 0.9% to 6.95% of dBR when changing the view separation from 50% to 100% of the coded baseline, respectively. Moreover, it was confirmed that the proposed UP-MVD enables up to 20% view separation adjustment while outperforming the anchor MVD format in all test sequences. Increasing the camera separation adjustment to 50%, still five out of seven sequences encoded with the proposed UP-MVD format outperformed the MVD anchor bitstreams in rate-distortion performance. As a future trend to further accomplish this research, the use of UP-MVD in multiview video

TABLE III. PERFORMANCE OF UP-MVD FORMAT WITH PROPOSED SCHEME 1 AGAINST ANCHOR

	Total (Coded PSNR)		Stereo pair (Base + 0.9)		Stereo pair (Base + 0.8)		Stereo pair (Base + 0.5)	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-5.90	0.22	-3.15	0.10	3.55	-0.14	32.59	-1.22
Poznan Street	-4.97	0.16	-5.37	0.18	-2.70	0.09	14.88	-0.49
Undo Dancer	-2.28	0.08	-4.05	0.15	-0.55	0.02	24.09	-0.74
Ghost Town Fly	-3.18	0.13	-5.22	0.20	-4.45	0.17	4.96	-0.19
Kendo	-14.10	0.79	-5.13	0.22	-0.98	-0.01	19.75	-1.05
Balloons	-9.06	0.50	-4.76	0.23	0.17	-0.05	20.52	-1.09
Newspaper	-9.14	0.42	-0.38	-0.01	12.22	-0.53	55.54	-1.80
<b>Average</b>	<b>-6.95</b>	<b>0.33</b>	<b>-4.01</b>	<b>0.15</b>	<b>1.04</b>	<b>-0.07</b>	<b>24.62</b>	<b>-0.94</b>

TABLE IV. PERFORMANCE OF UP-MVD FORMAT WITH PROPOSED SCHEME 2 AGAINST ANCHOR

	Total (Coded PSNR)		Stereo pair (Base + 0.9)		Stereo pair (Base + 0.8)		Stereo pair (Base + 0.5)	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-5.90	0.22	-5.51	0.18	-3.59	0.12	-16.07	0.60
Poznan Street	-4.97	0.16	-4.65	0.15	-4.18	0.14	-2.39	0.08
Undo Dancer	-2.28	0.08	-2.75	0.10	0.34	-0.01	11.08	-0.36
Ghost Town Fly	-3.18	0.13	-3.19	0.12	-2.96	0.11	-1.50	0.06
Kendo	-14.10	0.79	-11.58	0.59	-10.48	0.52	-7.26	0.33
Balloons	-9.06	0.50	-7.77	0.41	-6.58	0.34	-3.25	0.14
Newspaper	-9.14	0.42	-2.62	0.09	0.62	-0.06	13.09	-0.58
<b>Average</b>	<b>-6.95</b>	<b>0.33</b>	<b>-5.44</b>	<b>0.23</b>	<b>-3.83</b>	<b>0.16</b>	<b>-0.90</b>	<b>0.04</b>



with three or more views may be considered.

#### ACKNOWLEDGMENT

The authors would like to thank Professor M. Domański and his co-authors for providing Poznan sequences and their Camera Parameters.

#### REFERENCES

- [1] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "The zone of comfort: predicting visual discomfort with stereo displays," *Journal of Vision*, vol. 11, no. 8, July 2011
- [2] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," Jan. 2012
- [3] A. Smolic, K. Müller, P. Merkle, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, R. Tanger, P. Kauff, T. Wiegand, T. Balogh, Z. Megyesi, and A. Barsi, "Multi-view video plus depth (MVD) format for advanced 3D video systems," Joint Video Team, document JVT-W100, Apr. 2007
- [4] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proc. of SPIE stereoscopic displays and virtual reality systems XI*, pp. 93–104, Jan. 2004
- [5] "Call for proposals on 3D video coding technology," Moving Picture Experts Group, document N12036, Mar. 2011. Available: [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3dvcfp.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3dvcfp.zip)
- [6] [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3d-test-report.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3d-test-report.zip)
- [7] T. Suzuki, M. M. Hannuksela, Y. Chen, S. Hattori, and G. J. Sullivan (ed.), "MVC extension for inclusion of depth maps draft text 4," Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1001, July 2012
- [8] M. M. Hannuksela, Y. Chen, and T. Suzuki (ed.), "3D-AVC draft text 3," Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1002, Sep. 2012
- [9] D. Rusanovskyy and M. M. Hannuksela, "Description of 3D video coding technology proposal by Nokia," Moving Picture Experts Group, document M22552, Nov. 2011
- [10] C. Vazquez, W. J. Tam and F. Speranza, "Stereoscopic Imaging: Filling Disoccluded Areas in Depth Image-Based Rendering," *Proc. Of SPIE*, Vol. 6392, 2006.
- [11] P. Aflaki, M. M. Hannuksela, D. Rusanovskyy, and M. Gabbouj, "Non-Linear Depth Map Resampling for Depth-Enhanced 3D Video Coding," *IEEE Signal Processing Letters*, Vol. 20, issue 1, pp. 87-90, Jan. 2013
- [12] M. O. Wildeboer, T. Yendo, M. Panahpour Tehrani, T. Fujii, and M. Tanimoto, "Color based depth upsampling for depth compression," in *Proc. Picture Coding Symposium*, Dec. 2010, pp. 170–173.
- [13] K.-J. Oh, S. Yea, A. Vetro, and Y.-S. Ho, B, "Depth reconstruction filter and down/up sampling for depth coding in 3-D video," *IEEE Signal Process. Letters*, vol. 16, no. 9, pp. 747–750, Sep. 2009.
- [14] "Common test conditions for 3DV experimentation," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12560, Feb. 2012.
- [15] T. Leyvand, C. Meekhof, W. Yi-Chen Wei; Jian Sun; Baining Guo, "Kinect Identity: Technology and Experience," *Computer*, p.94-96, 2011
- [16] J. Zhu, L. Wang, R. Yang, J. Davis, Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps, *Proceedings of CVPR*, pp. 1-8, 2008.
- [17] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan multiview video test sequences and camera parameters," ISO/IEC JTC1/SC29/WG11 MPEG2009/M17050, Oct. 2009.

- [18] H.Schwarz, et al., "Description of 3D Video Technology Proposal by Fraunhofer HHI (MVC compatible)," ISO/IEC JTC1/SC29/WG11 MPEG2011/M22569, Nov, 2011.
- [19] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," ITU-T SG16 Q.6 document VCEG-M33, April 2001.

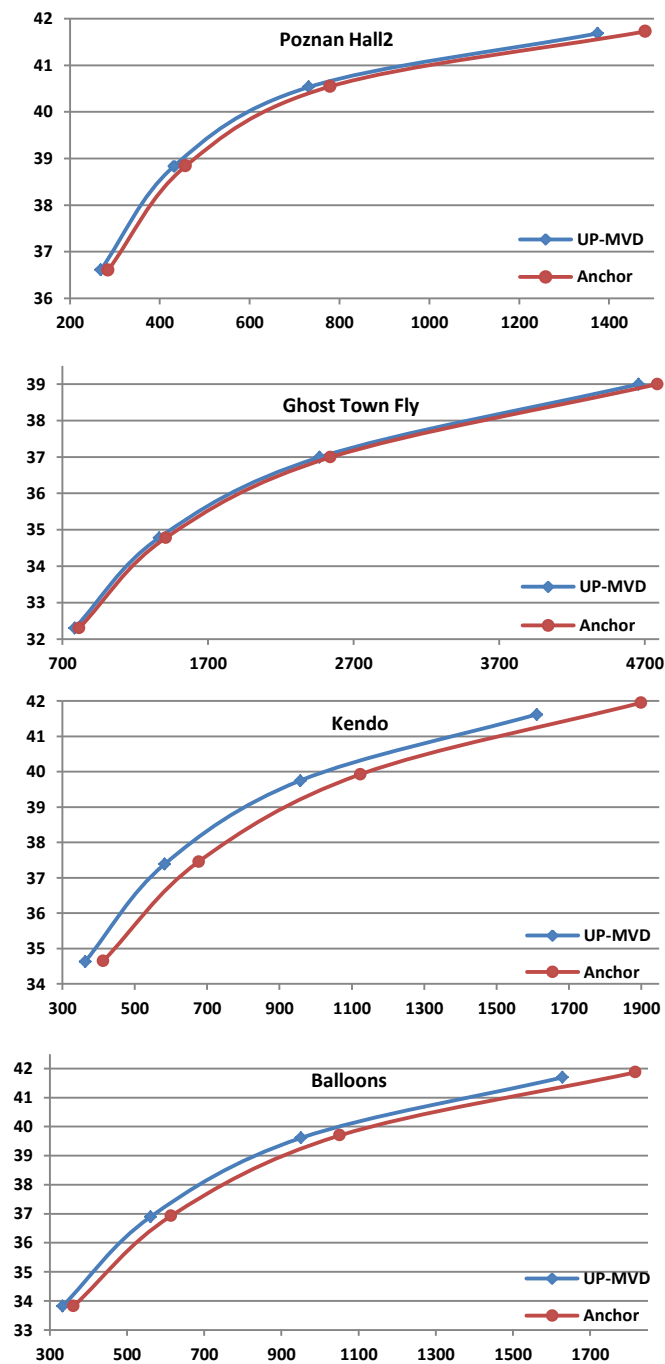


Figure 7. Rate distortion curves of four sequences with largest baseline stereopair (Base + 0.9) for proposed scheme 2

- [P5] **P. Aflaki**, Wenyi Su, Michal Joachimiak, D. Rusanovskyy, M. M. Hannuksela, Houqiang Li, and M. Gabbouj; “Coding of mixed-resolution multiview video in 3D video application, ” IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, September, 2013.

© IEEE, 2013, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publication\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publication_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# CODING OF MIXED-RESOLUTION MULTIVIEW VIDEO IN 3D VIDEO APPLICATION

*Payman Aflaki<sup>a</sup>, Wenyi Su<sup>b</sup>, Michal Joachimiak<sup>a</sup>, Dmytro Rusanovskyy<sup>c</sup>, Miska M. Hannuksela<sup>c</sup>,  
Houqiang Li<sup>b</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>University of Science and Technology of China, Hefei, China

<sup>c</sup>Nokia Research Center, Tampere, Finland;

## ABSTRACT

The emerging MVC+D standard specifies the coding of Multiview Video plus Depth (MVD) data for enabling advanced 3D video applications. MVC+D specifications define the coding of all views of MVD at equal spatial resolution and apply a conventional MVC technique for coding the multiview texture and the depth independently. This paper presents a modified MVC+D coding scheme, where only the base view is coded at the original resolution whereas dependent views are coded at reduced resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for dependent views. At the decoder side, the proposed scheme consists of a post-processing scheme which upsamples of the decoded views to their original resolution. The proposed scheme is compared against the original MVC+D scheme and an average of 4% delta bitrate reduction (dBR) in the coded views and 14.5% of dBR in the synthesized views are reported.

**Index Terms**— 3DV, MVC, asymmetric coding, spatial resolution, synthesized views

## 1. INTRODUCTION

The Moving Picture Experts Group (MPEG) has recently started 3D Video (3DV) standardization to enable support of advanced 3DV applications. The concept of advanced 3DV applications assumes that users can perceive a selected stereo-pair from numerous available views at the decoder side. Examples of such applications includes varying baseline to adjust the depth perception and multiview auto-stereoscopic displays (ASDs). Considering the complexity of capturing 3D scenes and the limitations in the distribution technologies, it is not possible to deliver a sufficiently large number of (20-50) views to the user's side with existing compression standards. To solve this problem, a 3D scene can be represented in multiview video plus depth (MVD) format [1] with a limited number of views, e.g. 2-3. The MVD data is coded and served as a source to a depth image-based rendering (DIBR) [2] algorithm which produces the required number of views at the decoder side.

In March 2011, MPEG issued a Call for Proposals for 3D video coding (hereafter referred to as the 3DV CfP) [3] for a new 3DV standard enabling the rendering of a selectable number of views with respect to the available

bitrate. As a result of the CfP evaluation [4], MPEG and, since July 2012, the Joint Collaborative Team on 3D Video Coding (JCT-3V) [5] have initiated development of a depth enhanced extension for MVC [6], abbreviated as MVC+D, to specify the encapsulation of coded MVD data into a single bitstream [7]. The MVC+D standard specifies MVD components (texture and depth) to have equal spatial resolution between different views and utilizes MVC technology [4] for the independent coding of texture and depth. As a result, a forward compatibility with MVC specification is preserved, and texture views of MVC+D bitstreams can be decoded with a conventional MVC decoder. The MVC+D specification was implemented in 3DV-ATM reference software [8] and was used in this study.

A possible solution to further reduce the bitrate and/or complexity of 3DV applications is to reduce the spatial resolution of a number of video views compared to the original resolution while preserving the original resolution for the remaining views. At the decoder side, views coded at the reduced resolution are upsampled to the original one using either conventional linear upsampling [9], or advanced super resolution techniques [10] that would benefit from multiview representation and the presence of depth. Being applied to texture component of MVD, this would result in a mixed-resolution texture representation and a significant bitrate reduction is hence expected.

It is obvious, that a scheme with a mixed-resolution texture representation would result in decoded views (e.g. stereoscopic image-pair) with different quality, which may affect stereoscopic perception. However, this argument can be addressed with the binocular rivalry theory [11] claiming that stereoscopic vision in the human visual system (HVS) fuses the images of an asymmetric quality stereoscopic image-pair so that the perceived quality is closer to that of the higher quality view. Several subjective quality evaluation studies have been conducted to investigate the use of the binocular rivalry theory in stereoscopic video coding [12-15]. Another work presented in [16] showed the applicability of asymmetric coding for MVC-like coding by encoding dependent views with a coarser quantization step compared to the base view. Subjective assessments confirmed that such coding scheme achieved a 20% bitrate reduction for stereoscopic image-pairs created from rendered views with no degradation in the perceived subjective quality.

This paper presents a modified MVC+D coding scheme, where only the base view is coded at the original resolution whereas dependent views are coded at a reduced resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for the inter-view predicted dependent views. At the decoder side, a post-processing scheme that performs upsampling of the decoded views back to their original resolution is proposed.

The rest of the paper is organized as follows. Section 2 presents the asymmetric texture coding schemes, while test material and simulation results are reported in Section 3. Finally, section 4 concludes the paper.

## 2. MVC CODING FOR MIXED-RESOLUTION TEXTURE REPRESENTATION

Let us assume that the 3DV system is coding MVD data representing a 3D scene with three viewing positions. In our description, we assume three-view (C3) coding scenario, since this is the most relevant test configuration with respect to the MPEG/JCT-3V Common Test Condition [17].

The flowchart of the proposed 3DV system with a mixed-resolution texture representation is shown in Figure 1. An arbitrary view (e.g. the center view) of the input MVD data is coded with H.264/AVC at the original resolution. According to H.264/MVC specification, this view is considered as a base view and provides reference pictures for the inter-view prediction and the coding of dependent views. In the proposed scheme, dependent views of MVD data are coded at a reduced resolution, thus the proposed scheme downsamples the data at the pre-processing stage and upsamples it back to original resolution at the post-processing stage, as shown in Figure 1.

In this study, the base view was coded at the original full resolution (FR) whereas dependent views were coded at

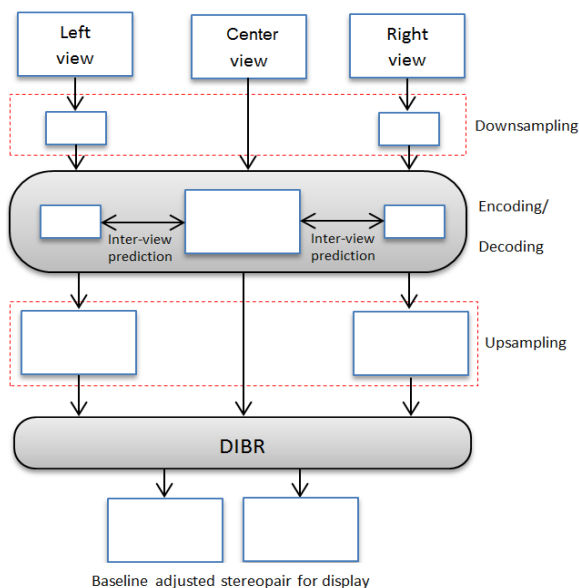


Figure 1. Block diagram of the proposed encoding scheme

half of the original resolution along each direction, which resulted in quarter resolution (QR) downsampled view. However, the downsampling ratio can be adjusted based on the target application.

Figure 2 shows a simplified flowchart of H.264/MVC scheme with the proposed modification in the in-loop operations for enabling mixed-resolution coding. Base view coding is performed with the conventional H.264/AVC technique and the decoded pictures are stored in a frame buffer. Since the resolution of the base view is different from that of the dependent view, the decoded picture of the original view cannot be used as a reference picture for coding dependent views. To enable inter-view prediction, the resolution of the base view picture should match the resolution of dependent views. There are various approaches to do this, and in this paper we tested two methods: decimation of the reference picture (marked with green line in Figure 2) and downsampling of the decoded picture (marked with blue line). The following sections present the motivation and describe the proposed schemes in details.

### 2.1 Low complexity Coding (Scheme 1)

The specification of H.264/MVC defines Motion Compensating Prediction (MCP) with quarter-pixel (Q-pel) resolution of motion vectors. To achieve this, in each decoded image view, which is marked to be used as a reference, undergoes in-loop interpolation by a factor of 4 in the horizontal and vertical directions. The interpolated picture is stored in a frame buffer of the corresponding view and used as a reference picture for inter-prediction (temporal MCP). In addition, the reference picture of the base view can be used as a reference for inter-view prediction when a dependent view is coded. However, in the case of mixed-resolution coding, the reference picture produced in the base view is 2x larger than the reference pictures produced in dependent views and hence cannot be used in the same MCP. To solve this problem, the inter-view reference picture (Q-pel resolution) of the base view is decimated by a factor of 2 along each direction and the subsampled version is placed in the reference frame buffer of the dependent view, shown by the green line module in Figure 2.

The algorithm proposed in this section (scheme 1) has a negligible complexity increase and introduces minimal changes to the H.264/MVC architecture. It is believed that such changes can be performed by software only update to the already deployed decoding infrastructure.

However, this algorithm does not take into consideration parameters of downsampling performed to the dependent view at the post-processing stage, e.g. the low pass filter (LPF) phase, and its performance may suffer from a possible mismatch in the pixel location grid used in the base and dependent views, and aliasing, since the decimation procedure does not apply any LPF. This may lead to sub-optimal performance of the MCP in the inter-view prediction.

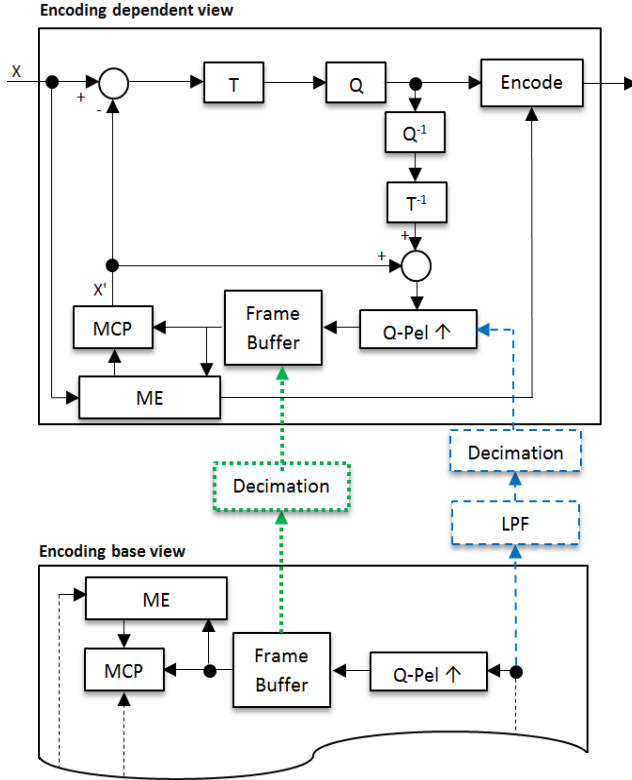


Figure 2. MVC coding for mixed-resolution video, where the proposed Scheme 1 is depicted in green box and Scheme 2 in blue

## 2.2 High performance Coding (Scheme 2)

To overcome the problem raised in the previous sub-section, the reference picture of the base view which is to be used for inter-view prediction should be downsampled with a proper antialiasing low pass filtering applied prior to decimation. The decoded picture of the base view is downsampled and undergoes Q-pel interpolation in the dependent view, and thus it is handled independently from the MCP chain of the base view. The proposed alternative solution is shown in Figure 2 with processing modules marked with blue dashed lines.

It is essential for in-loop downsampling applied to pictures of the base view to use an identical filter as the one used in the preprocessing of the dependent views. This will require adequate signaling in the Sequence Parameters Set; however, it will ensure an identical pixel location grid for the dependent view and the reference picture of the base view.

The algorithm proposed in this section (scheme 2) has a larger computational complexity in comparison to scheme 1, since it performs antialiasing low pass filtering and additional Q-Pel interpolation. However, the absence of aliasing artifacts along with no mismatch in pixel location grid between the coded and the reference images are expected to contribute towards efficient inter-view

prediction. The simulation results provided in the next section confirm these expectations.

## 3. TEST MATERIAL AND SIMULATION RESULTS

Both schemes proposed in this paper (Scheme 1 and Scheme 2) were integrated to the 3DV-ATM software and compared against the anchor scheme (MVC+D). Simulations were conducted under the specifications of C3 scenario of 3DV Common Test Condition (CTC) [17] and JCT-3V/MPEG MVD test sequences were utilized. In this scenario three depth-enhanced texture views are encoded and then several possible in-between views are synthesized to be exploited in stereoscopic image-pair creation.

The full resolution MVC+D coding, as implemented in 3DV-ATM [8], and 3DV VSRS [18] were utilized to produce a full resolution anchor results. Table I summarizes the major parameters used for the 3DV-ATM configuration, whereas complete configuration files for MVC+D are available in [17].

The simulation framework for the proposed schemes (Scheme 1 and Scheme 2) is specified as shown in Table I and the following changes were introduced.

The following pre-processing and post-processing stages as shown in Figure 1 were utilized to produce simulation results for the proposed Scheme 1 and Scheme 2.

### Pre-processing:

Texture views of MVD data marked to be coded as dependent views were downsampled at the pre-processing stage. The downsampling was performed with a lowpass filter used in [19]. The LPF is designed with a cut-off frequency of  $0.9\pi$  and has 12 filter taps. The filter coefficients are as follows:

$$h1 = [2 -3 -9 6 39 58 39 6 -9 -3 2 0]/128 \quad (1)$$

### Post-processing:

Following the decoding and prior to the DIBR, the decoded dependent views were upsampled by a factor of 2 in the horizontal and vertical directions back to the original resolution. The upsampling was performed with the 6-tap H.264/AVC interpolation filter [9]. The coefficients of this

TABLE I. CONFIGURATION OF 3DV-ATM CONFIGURED THE ANCHOR (MVC+D) AND PROPOSED SCHEME

Coding Parameters	Settings
Compatibility Mode	0 (MVC+D)
Multi-view scenario	Three views (C3)
MVD resolution ratio (Texture : Depth)	1:0.5
Inter-view prediction structure	PIP
Inter prediction structure	HierarchicalB, GOP8
QP settings for texture & depth	26, 31, 36, 41
Encoder settings	RDO ON, VSO OFF
View Synthesis in Post-processing	Fast_1D VSRS [18]
Test sequences and coded, synthesized views	As specified in [17]

filter are as follows:

$$h2 = [1 \ -5 \ 20 \ 20 \ -5 \ 1]/32 \quad (2)$$

**Proposed schemes:**

Integration of Scheme 1 to the 3DV-ATM software was straightforward and its details were given in sub-section 2.1. Scheme 2 as described in Section 2.2 was integrated to 3DV-ATM and the filter given in equation (1) was used.

The compression efficiency of the proposed schemes was evaluated according to the CTC [17] specification. The Bjontegaard delta bitrate and delta Peak Signal-to-Noise Ratio (PSNR) metrics [20] were utilized for these purposes and the MVC+D scheme was used as the anchor. The delta bitrate reduction (dBR) is presented for the total coded views (the total bitrate of the texture and depth coding along with PSNR of the texture views) and the synthesized views (the total bitrate of the texture and depth coding along with the PSNR of the synthesized views). The PSNR of the synthesized views at the decoder side were computed against the reference view synthesis results, as specified in CTC [17] and achieved from the original uncompressed texture and depth information. The results comparing the proposed schemes against the MVC+D anchor are reported in Tables II and III. Moreover, rate-distortion (RD) curves achieved with Scheme 2 and for the synthesized views of Poznan Hall 2 sequence are depicted in Figure 3. These curves well match with dBR values presented in Table III, confirming higher efficiency of Scheme 2 against anchor.

As reported in Tables II and III, both proposed MVC+D schemes with mixed-resolution texture representation outperformed the full resolution MVC+D anchor. The low complexity Scheme 1 reduces the average coded bit rate by

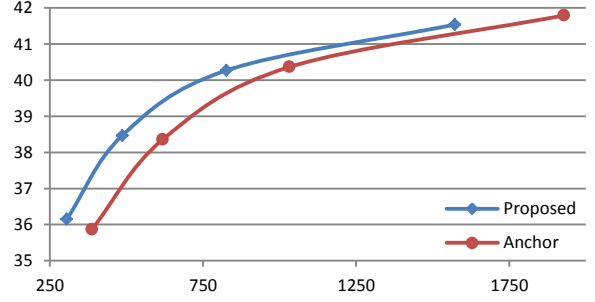


Figure 3. Rate-distortion curve for synthesized views for the sequence Poznan Hall used in Scheme 2 against the anchor

1.35% compared to the anchor, whereas the average compression gain for all natural sequences (excluding Ghost Town Fly and Dancer) is more than 10% of dBR. For the synthesized views, Scheme 1 provides 12.63% dBR on average (synthetic sequences included). A possible explanation for this effect is the fact that DIBR operates at the low resolution depth map, see Table I, and therefore, rendering becomes less accurate and high frequency components of synthetic sequences may not bias the final PSNR of synthesized views.

The high performance Scheme 2, as expected, provides a larger coding gain, outperforming the MVC+D anchor by 4.06% of dBR on average for coded bitrates and by 14.52% of dBR on average for synthesized views. It should be noted that Scheme 2 significantly outperforms Scheme 1 for synthetic sequences, where the impact of aliasing artifacts and mismatch in pixel grid seem to degrade inter-view prediction in Scheme 1. On the other hand, coding performance for natural sequences seems to be very close for both Scheme 1 and Scheme 2, giving about 10% of dBR for coded views and about 14% of dBR gain for synthesized views against the MVD anchor, respectively.

**4. CONCLUSIONS**

The paper proposed a novel modified MVC+D coding scheme that supports the coding MVD data with a mixed-resolution texture representation. We proposed to encode only the base view at the original resolution whereas the spatial resolution of dependent views is reduced. At the decoder side, the proposed scheme consists of a post-processing scheme that performs upsampling of decoded views back to their original resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for dependent views. The proposed scheme was compared against the original MVC+D and objective coding gains of 4% of average delta bitrate reduction (dBR) and 14.5% of dBR on synthesized views were reported.

**5. ACKNOWLEDGEMENT**

The authors would like to thank M. Domański, et al. for providing Poznan sequences and Camera Parameters [21].

TABLE II. PERFORMANCE OF THE PROPOSED MIXED-RESOLUTION SCHEME 1 COMPARED TO THE ANCHOR

	Coded views		Synthesized views	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-18.12	0.60	-20.22	0.75
Poznan Street	-2.16	0.00	-8.96	0.27
Undo Dancer	30.74	-1.22	-12.39	0.32
Ghost Town Fly	10.47	-0.83	-6.43	0.11
Kendo	-12.14	0.59	-14.46	0.69
Balloons	-13.35	0.68	-15.47	0.77
Newspaper	-4.90	0.17	-10.45	0.39
<b>Average</b>	<b>-1.35</b>	<b>0.00</b>	<b>-12.63</b>	<b>0.47</b>

TABLE III. PERFORMANCE OF THE PROPOSED MIXED-RESOLUTION SCHEME 2 COMPARED TO THE ANCHOR

	Coded views		Synthesized views	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-18.29	0.62	-20.58	0.78
Poznan Street	0.04	-0.09	-8.39	0.25
Undo Dancer	18.98	-0.88	-18.27	0.54
Ghost Town Fly	1.58	-0.41	-12.96	0.40
Kendo	-12.36	0.60	-14.88	0.71
Balloons	-13.44	0.68	-15.86	0.79
Newspaper	-4.90	0.15	-10.71	0.40
<b>Average</b>	<b>-4.06</b>	<b>0.10</b>	<b>-14.52</b>	<b>0.55</b>

## 6. REFERENCES

- [1] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," Proc. of IEEE International Conference on Image Processing, vol. 1, pp. 201-204, Oct. 2007.
- [2] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, CA, U.S.A., Jan. 2004, pp. 93-104.
- [3] MPEG Video and Requirement Groups, "Call for Proposals on 3D Video Coding Technology", MPEG output document N12036, Geneva, Switzerland, March 2011
- [4] [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3d-test-report.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3d-test-report.zip)
- [5] T. Suzuki, M. M. Hannuksela, Y. Chen, S. Hattori, and G. J. Sullivan (ed.), "MVC extension for inclusion of depth maps draft text 4," Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1001, July 2012
- [6] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.
- [7] M. M. Hannuksela, Y. Chen, and T. Suzuki (ed.), "3D-AVC draft text 3," Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1002, Sep. 2012
- [8] "Test model for AVC based 3D video coding," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12558, Feb. 2012
- [9] JSVM Software  
[http://ip.hhi.de/imagecom\\_G1/savce/downloads/SVC-Reference-Software.htm](http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm)
- [10] P. Milanfar, Ed., "Super-Resolution Imaging". Boca Raton, FL: CRC Press, 2010.
- [11] R. Blake, "Threshold conditions for binocular rivalry," Journal of Experimental Psychology: Human Perception and Performance, vol. 3(2), pp. 251-257, 2001.
- [12] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, "Subjective Study on Compressed Asymmetric Stereoscopic Video," Proc. of IEEE Int. Conf. on Image Processing (ICIP), Sep. 2010.
- [13] W. J. Tam, "Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV," Joint Video Team document JVTW094, Apr. 2007.
- [14] P. Seuntjens, L. Meesters, and W. IJsselstein, "Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation," ACM Transactions on Applied Perception, vol. 3, no. 2, pp. 95-109, Apr. 2006.
- [15] H. Brust, A. Smolic, K. Müller, G. Tech, and T. Wiegand, "Mixed-resolution coding of stereoscopic video for mobile devices" 3DTV Conference, May 2009.
- [16] P. Aflaki, D. Rusanovskyy, T. Utriainen, E. Pesonen, M. M. Hannuksela, S. Jumisko-Pyykkö, and M. Gabbouj, "Study of asymmetric quality between coded views in depth-enhanced multiview video coding," International Conference on 3D Imaging (IC3D), Dec. 2011.
- [17] "Common test conditions for 3DV experimentation," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12560, Feb. 2012.
- [18] H. Schwarz, et al., "Description of 3D Video Technology Proposal by Fraunhofer HHI (MVC compatible)," ISO/IEC JTC1/SC29/WG11 MPEG2011/M22569, Nov, 2011.
- [19] J. Dong, Y. He, Y. Ye, "Downsampling filters for anchor generation for scalable extensions of HEVC," ISO/IEC JTC1/SC29/WG11 MPEG2012/M23485, Feb. 2012.
- [20] G. Bjøntegaard, "Calculation of average PSNR differences between RD-Curves," ITU-T SG16 Q.6 document VCEG-M33, April 2001
- [21] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, Xian, China, October 2009.

- [P6] **P. Aflaki**, M. M. Hannuksela, M. Homayouni, and M. Gabbouj; “Cross-asymmetric mixed-resolution 3D video compression, ” International 3DTV CONF, Zurich, Switzerland, October, 2012.

© IEEE, 2012, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.



# CROSS-ASYMMETRIC MIXED-RESOLUTION 3D VIDEO COMPRESSION

*Payman Aflaki<sup>a</sup>, Miska M. Hannuksela<sup>b</sup>, Maryam Homayouni<sup>a</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

## ABSTRACT

Conventional mixed-resolution (MR) stereoscopic video where one view has full resolution (FR) and the other view has a lower resolution has shown to provide similar subjective quality compared to symmetric FR stereoscopic video while decreasing the encoding complexity considerably. In this paper, we propose a new cross-asymmetric mixed-resolution scheme where both views have a lower resolution compared to FR but downsampling is applied in horizontal direction for one view while the other view is vertically downsampled. Subjective results comparing the proposed scheme with the conventional MR and the symmetric FR schemes show that the perceived quality of the proposed scheme is higher than that of the two other schemes. Moreover, the computational complexity and memory requirements also are reduced thanks to the decreased number of pixels involved in the encoding and decoding processes.

**Index Terms** — Asymmetric stereoscopic video, mixed resolution, subjective evaluation

## 1. INTRODUCTION

Two approaches for compressing stereoscopic video are common nowadays: frame-compatible stereoscopic video and Multiview Video Coding (MVC) [1]. The latter was standardized as an annex to Advanced Video Coding (H.264/AVC) standard [2]. In frame-compatible stereoscopic video, a spatial packing of a stereo pair into a single frame is performed at the encoder side as a pre-processing step for encoding and then the frame-packed frames are encoded with a conventional 2D video coding scheme. The encoder side may indicate the used frame packing format for example by including one or more frame packing arrangement supplemental enhancement information (SEI) messages as specified in the H.264/AVC standard into the bitstream. The decoder unpacks the two constituent frames from the output frames of the decoder and upsamples them to revert the encoder side's downsampling process and render the constituent frames on a 3D display. In contrast to frame packing, MVC enables any spatial resolution to be used in encoding and facilitates plain H.264/AVC decoders to produce single-view output without additional processing. Moreover, inter-view prediction presented in MVC provides a considerable compression improvement compared to frame packing and stereoscopic video representation using H.264/AVC simulcast. However, due to increased amount of data compared to conventional 2D video, further compression without perceivable subjective quality degradation is required in many applications.

One potential approach to achieve a better compression is to provide left and right views with different qualities referred to as asymmetric quality video where one of the two views is coded with a lower quality compared to the other one. This is attributed to the widely believed assumption of the binocular suppression theory [3] that the Human Visual System (HVS) fuses the two images such that the perceived quality is close to that of the higher quality view. Quality difference can be achieved by utilizing coarser quantization steps for one view and/or presenting stereoscopic video with MR where one view is downsampled prior to encoding. Considering that a smaller number of samples are involved in the coding/decoding process of MR stereoscopic video, it is expected to have lower processing complexity compared to the FR scheme.

Asymmetric stereoscopic video coding has been studied extensively over the years. For example, in [4] a set of subjective tests on encoded FR and MR stereoscopic videos were performed under the same bitrate constraint. The results show that the MR stereoscopic video with downsampling ratio 1/2, applied both vertically and horizontally, performed similarly to the FR in most cases. In [5], the MR approach was compared with a quality-asymmetric approach, in which the bigger steps were utilized for transform coefficients while coding one of the views. Results confirmed that perceived quality of the MR videos were close to that of the FR view. The impact of quantization was verified in [6], which concluded that the perceived quality of coded equal-resolution stereo image pairs was approximately the average of the perceived qualities of the high-quality image and the low-quality image of the stereo pairs.

To approximate the perceived quality of the stereoscopic video, objective quality metrics often perform well. However, in the case of asymmetric stereoscopic video, there are two views with different qualities, and it has been found that objective quality assessment metrics face some ambiguity on how to approximate the perceived quality of asymmetric stereoscopic video [7].

This paper first describes a new cross-asymmetric MR compression technique and then evaluates its performance with a set of subjective tests. The proposed method is compared to compressed FR and conventional MR videos with different downsampling ratios applied to one view and along both coordinate axes. JM 17.2 reference software [8] of H.264/AVC is utilized as the encoder and the comparison is performed under the same bitrate constraints for two different bitrates.

This paper is organized as follows. In Section 2, the proposed MR scheme is presented while the test material and procedure are explained in section 3. Section 4 presents and discusses the results, and section 5 concludes the paper.

## 2. PROPOSED CROSS-ASYMMETRIC MIXED-RESOLUTION SCHEME

### 2.1 Overview

The traditional MR scheme performs downsampling on one view while the other view remains untouched having FR. In order to apply inter-view prediction between views of different resolution, a resampling process is required in the coding and decoding loop. As no such resampling is available in H.264/MVC, we present the proposed scheme in the context of H.264/AVC simulcast. Consequently, we also avoid any influence of non-standardized resampling and inter-view prediction algorithms on the results. Hence, we can be sure that the results are trustable and different performances are only due to utilization of different MR schemes, rather than different performance of utilized MR adaptive H.264/MVC codec. While the proposed method is presented for H.264/AVC simulcast environment, a frame packing scheme can also be designed, or a multiview codec with in-loop resampling and inter-view prediction can be applied.

### 2.2 Proposed MR scheme

In the proposed cross-asymmetric MR scheme, different resolutions for left and right views are utilized. Unlike the conventional MR scheme, we intend not to utilize any of two views in FR but downsample both views asymmetrically in such a way that horizontal and vertical downsampling ratios differ for the same view and the choice of the horizontal and vertical downsampling ratios is reversed for the other view. In other words, one view is downsampled more in the vertical direction while keeping more horizontal spatial information in that view. The other view is downsampled more in the horizontal direction. On the basis of the binocular suppression theory we expect the human visual system to perceive the picture in such a way that the higher quality information in each direction from the view where less downsampling along that direction was applied prevails. Figure 2 presents the general block diagram of the proposed MR scheme.  $W$  and  $H$  represent the width and height of the FR views, respectively, while  $a_1$ ,  $a_2$ ,  $b_1$ , and  $b_2$  are downsampling coefficients. In the proposed scheme, it is required that  $a_1 > a_2$  and  $b_1 < b_2$ . Since the encoding is applied on downsampled views, after decoding and prior to the final presentation of stereoscopic video, the views will be upsampled to the FR.

Considering that eye dominance was shown not to have an impact on the perceived quality of MR stereoscopic videos [9], it is proposed that the decision on which the view should be more downsampled in horizontal/vertical direction is made based on spatial information (SI) [10] along each direction of each view. To calculate SI, a  $3 \times 3$  Sobel filter is utilized (1) to emphasize horizontal edges using the smoothing effect by approximating a vertical gradient. To emphasize vertical edges, the transpose of the filter ( $h'$ ) will be applied.

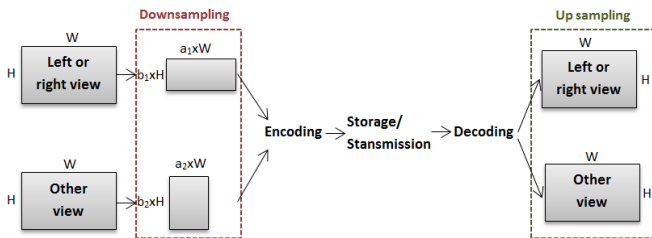


Figure 2. Proposed MR encoding/decoding scheme

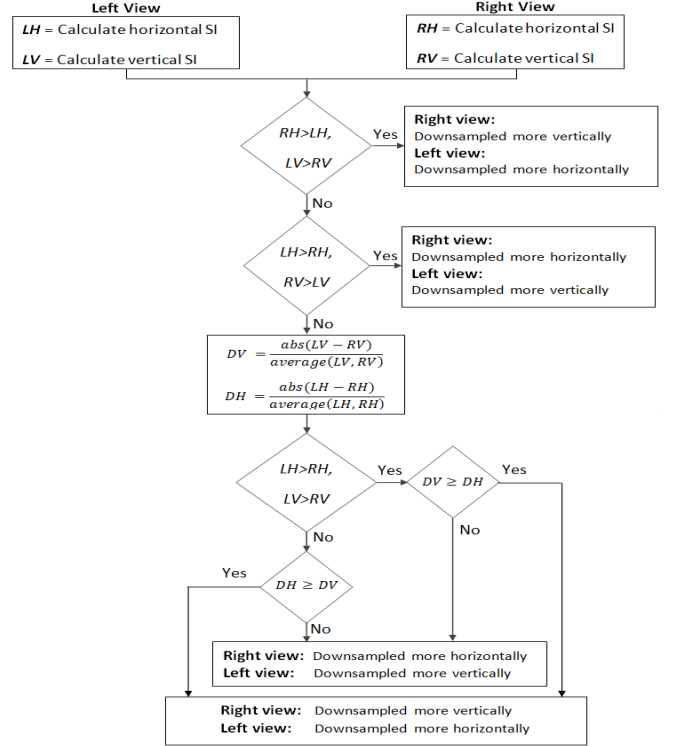


Figure 3. Flowchart of selecting the downsampling direction for left and right view.

$$h = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (1)$$

Based on the direction of the applied Sobel filter to the luma values of each image, SI will be calculated for the vertical or horizontal direction averaging over the magnitudes of the filtered image. Considering  $LV$ ,  $LH$ ,  $RV$ , and  $RH$  presenting SI of Left view in Vertical direction, Left view in Horizontal direction, Right view in Vertical direction, and Right view in Horizontal direction, respectively. The flowchart presented in Figure 3 shows how the decision of the downsampling direction of the left and right views is made. If  $LV$  is greater than  $RV$  and  $LH$  is smaller than  $RH$ , then the right view will be downsampled more in the vertical direction and the left view will be downsampled more in the horizontal direction. On the other hand, if  $LV$  is smaller than  $RV$  and  $LH$  is greater than  $RH$ , then the right view will be downsampled more in the horizontal direction while the left view is downsampled more in the vertical direction. If none of the above mentioned cases is valid, i.e. the left view has a higher SI in both directions ( $LV > RV$  and  $LH > RH$ ) or the right view has a higher SI in both directions ( $LV < RV$  and  $LH < RH$ ), the decision is made based on the normalized absolute difference levels, defined next. Considering the case where the left view has a higher SI in both directions, let us define the normalized absolute difference values as:

$$DV = \frac{abs(LV - RV)}{average(LV, RV)}$$

$$DH = \frac{abs(LH - RH)}{average(LH, RH)} \quad (2)$$

where  $average(LX, RX) = \frac{LX + RX}{2}$  and  $DV$  and  $DH$  present the normalized absolute difference between the left and right views in vertical and horizontal direction, respectively. If  $DV \geq DH$ , then

the left view will be downsampled more in the horizontal direction and the right view will be downsampled more in the vertical direction. In the case where the right view has a higher SI in both directions, if  $DV \geq DH$  then the right view will be downsampled more in the horizontal direction and the left view will be downsampled more in the vertical direction. The main idea behind the use of SI for downsampling the left and right views is that the highest combined amount of information in the downsampled MR stereoscopic video is preserved.

### 3. TEST SETUP

#### 3.1 Test material

The tests were carried out using four sequences: Ballet, Breakdancer [11], Alt Moabit, and Book Arrival [12] with resolution 1024×768.

Five types of encoding schemes based on the resolution of left and right views were selected for the subjective test:

1. Anchor scheme: Full-resolution in both views (AS)
2. Conventional MR Scheme with downsampling ratio = 1/2, i.e. half resolution for one view in both directions and FR in the other view (CS1/2)
3. Proposed MR Scheme with downsampling ratio = 1/2, i.e. one view is downsampled only in vertical direction while the other view is only downsampled in the horizontal direction, the downsampling ratio is set to 1/2 for both cases (PS1/2)
4. Conventional MR Scheme with downsampling ratio = 1/4, (CS1/4)
5. Proposed MR Scheme with downsampling ratio = 1/4, (PS1/4)

The filters included in the JSVM reference software of the Scalable Video Coding standard were utilized in the downsampling and upsampling operations [13]. Moreover, views were independently coded using the reference JM 17.2 software in order to treat the FR and MR cases as equally as possible, as described in sub-section 2.1.

The quality and bitrate of H.264/AVC bitstreams is controlled by the quantization parameter (QP). In order to get results from a larger range of qualities and compressed bitrates, two constant QP values, 34 and 38, were selected for encoding in AS. Other schemes were encoded having a bitrate within 4% of the bitrate of the corresponding AS. The QP for left and right view was selected in such a way that bitrate ratio between the left and right view was close to one. This was due to the fact that we did not

want to affect the experiment by the selection of different QPs but limit the study to evaluate the performance of different applied downsampling schemes. The uncompressed FR sequences were included in the viewed sequences to obtain a reference point for the highest perceived quality of each particular sequence.

#### 3.2 Test procedure

Test clips were displayed on a 46" polarizing stereoscopic screen having a total resolution of 1920×1200 pixels and a resolution of 1920×600 per view when used in the stereoscopic mode. Sequences were presented un-scaled with black background on the display fixing the viewing distance to 1.63 meter that is 4 times the height of the videos.

The duration of a viewing session was limited to ~35 minutes to avoid viewers becoming exhausted. In total 20 subjects (17 male and 3 female) attended the test. The average age of subjects was 26.5 years. All the participants were naïve users who had no previous experience on 3D video processing.

Subjective quality assessment was done according to Double Stimulus Impairment Scale (DSIS) method [14] and discrete unlabeled quality scale from 0 to 10 was used for quality assessment. Prior to the actual test, subjects were familiarized with test task, test sequences and with the variation in quality they could expect in the actual tests. The viewers were instructed that 0 stands for the lowest quality and 10 for the highest. The test clips were presented in a random order each clip was rated independently after its presentation. Prior to the participation in subjective viewing experiment, candidates were subject to a thorough vision screening. All participants had a stereoscopic acuity of at least 60 arc sec.

### 4. RESULTS AND DISCUSSION

The average and 95% confidence interval (CI) of subjective scores are presented in Figure 4. The naming of the encoding schemes is according to sub-section 3.1 and O represents the original FR uncompressed stereoscopic video.

It can be judged from the mean scores and confidence intervals presented in Figure 4 that the subjective quality of the higher bitrate was rated better in general compared to the lower bitrate. Moreover, the original uncompressed video had superior quality compared to other schemes. The observation on significant differences between the encoding schemes was further analyzed using statistical analysis as presented in the paragraphs below.

Non-parametric statistical analysis methods, Friedman's and

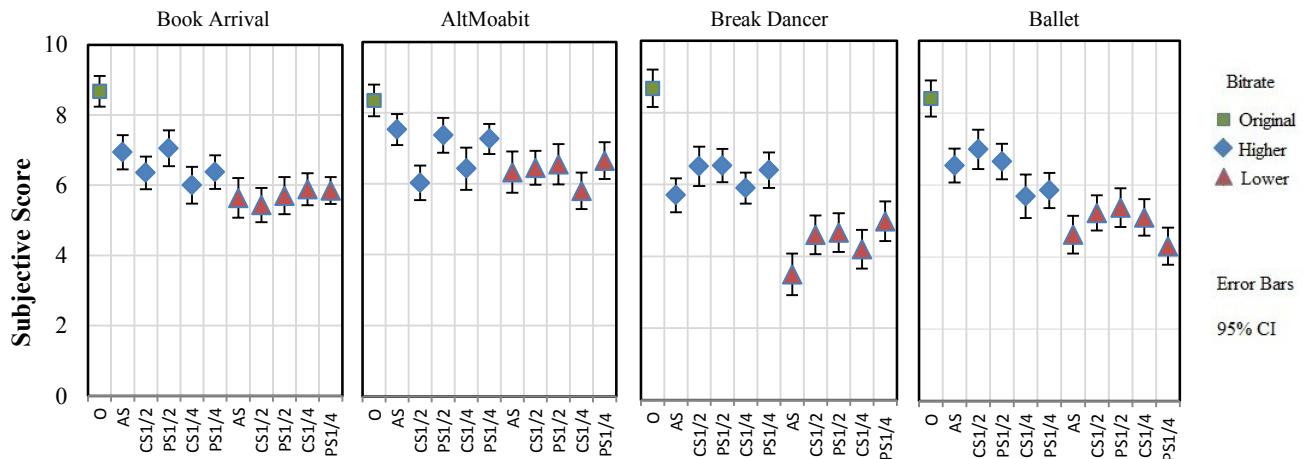


Figure 4. Viewing experience ratings

Table 1. Pairwise performance comparison of different coding schemes over all content

Coding scheme	Higher bitrate			Lower bitrate		
	<i>Better</i>	<i>Similar</i>	<i>Worse</i>	<i>Better</i>	<i>Similar</i>	<i>Worse</i>
AS	5	8	3	0	12	4
CS1/2	3	9	4	2	14	0
PS1/2	9	7	0	4	12	0
CS1/4	0	7	9	1	12	3
PS1/4	3	9	4	3	10	3

Table 2. Per view and total number of pixels involved in the coding/decoding process for different coding schemes

Coding Scheme	Number of Pixels		
	One view	Other view	Total
AS	$W * H$	$W * H$	$2 * W * H$
CS1/2	$\frac{W}{2} * \frac{H}{2}$	$W * H$	$\frac{5}{4} * W * H$
PS1/2	$\frac{W}{2} * H$	$W * \frac{H}{2}$	$W * H$
CS1/4	$\frac{W}{4} * \frac{H}{4}$	$W * H$	$\frac{17}{16} * W * H$
PS1/4	$\frac{W}{4} * H$	$W * \frac{H}{4}$	$\frac{W * H}{2}$

Wilcoxon's tests, were used as the data did not reach normal distribution (Kolmogorov-Smirnov:  $p < .05$ ). Wilcoxon's test is applicable to measure differences between two related and ordinal data sets [15]. A significance difference level of  $p < 0.05$  was used in our analysis.

Table 1 reports the performance analysis results of each coding scheme, as achieved by Wilcoxon's test, in a pairwise comparison to other schemes. For each coding scheme three values are reported per bitrate. First, the value in column *Better* provides the total number of cases in which the associated scheme was ranked significantly better than the other schemes. The second number reports the total number of cases where similar subjective quality to the other schemes was reported (*Similar*). Finally, the third value in column *Worse*, reports the number of cases in which the referred coding scheme provided a significantly worse rating compared to the other schemes. The next paragraphs discuss the performance of different coding schemes based on the statistics reported in Table 1.

In the higher bitrate the performance of PS1/2 was clearly superior, since in no comparison it was ranked worse than other schemes and in the majority of cases it was ranked better compared to other schemes. Moreover, CS1/4 performed worse since in the majority of comparisons it was ranked worse than other schemes while in none of the comparisons it was ranked better.

In the lower bitrate, the coding schemes performed closer to each other while the majority of comparisons resulted in a similar subjective quality. AS performed slightly inferior to others since it was never ranked better. Moreover, PS1/2 and CS1/2 were never ranked worse compared to other schemes; nevertheless, PS1/2 had slightly better performance compared to CS1/2 since it was ranked better in four cases compared to two cases for CS1/2.

In general, the results show that utilization of FR videos (as in AS) in that lower bitrate was not subjectively preferred and applying downsampling through different schemes provided a higher perceived quality. This is in agreement with the conclusion achieved in [4, 16]. Moreover, the subjective results confirm that PS1/2 performed the best in the lower and higher bitrates.

Next we compare the complexity of the coding/decoding schemes based on the number of pixels involved in the

coding/decoding process. If the width and the height of the FR views is represented with  $W$  and  $H$ , respectively, the total number of pixels for both views can be calculated as shown in Table 2. Based on the results presented in Table 2 the proposed methods (PS1/2 and PS1/4) introduce the least number of pixels for the coding/decoding process. Hence, along with superior subjective quality, lower complexity is another important advantage which justifies the utilization of the proposed coding scheme.

## 5. CONCLUSIONS

The paper proposes a mixed-resolution (MR) stereoscopic video coding scheme, where one view is horizontally downsampled while the other view is vertically downsampled at different rates. The proposed scheme was compared with symmetric full-resolution (FR) stereoscopic video as well as the conventional MR coding, where one view is downsampled along both coordinate axes while the other view is maintained at its original resolution. A series of subjective tests was conducted comparing the proposed scheme with conventional MR and symmetric FR schemes. The results show that proposed method outperforms the other methods while decreasing the computational complexity and memory requirements of the codec.

## 6. REFERENCES

- [1] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," EURASIP Journal on Advances in Signal Processing, vol. 2009
- [2] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," Mar. 2009.
- [3] R. Blake, "Threshold conditions for binocular rivalry," Journal of Experimental Psychology: Human Perception and Performance, vol. 3(2), pp. 251-257, 2001.
- [4] P. Aflaki, et al, "Subjective study on compressed asymmetric stereoscopic video," Proc. of Int. Conf. on Image Proc., Sep. 2010.
- [5] W. J. Tam, "Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV," Joint Video Team document JVT-W094, Apr. 2007.
- [6] P. Seuntjens, L. Meesters, and W. IJsselstein, "Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation," ACM Trans. on Applied Perception, vol. 3, no. 2, pp. 95-109, Apr. 2006.
- [7] P. W. Gorley, N.S. Holliman, "Stereoscopic image quality metrics and compression", Stereoscopic Displays and Virtual Reality Systems XIX, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 6803, January 2008
- [8] JM reference software: <http://iphome.hhi.de/suehring/tml/download>
- [9] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video," Proc. of 3DTV-Conference, Jun. 2010.
- [10] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 1999.
- [11] <http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload>
- [12] <ftp://ftp.hhi.de/HHIMPEG3DV/sequences/>
- [13] JSVM Software [http://ip.hhi.de/imagecom\\_G1/save/downloads/SVC-Reference-Software.htm](http://ip.hhi.de/imagecom_G1/save/downloads/SVC-Reference-Software.htm)
- [14] ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002
- [15] H. Cooligan "Research methods and statistics in psychology," (4th ed.). London: Arrowsmith., 2004.
- [16] H. Brust, A. Smolic, K. Müller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices" 3DTV Conference, May 2009.

- [P7] **P. Aflaki**, D. Rusanovskyy, M. M. Hannuksela, and M. Gabbouj; “Frequency based adaptive spatial resolution selection for 3D video coding,” European Signal Processing Conference (EUSIPCO), Bucharest, Romania, August, 2012.

© IEEE, 2012, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# FREQUENCY BASED ADAPTIVE SPATIAL RESOLUTION SELECTION FOR 3D VIDEO CODING

*Payman Aflaki<sup>a</sup>, Dmytro Rusanovskyy<sup>b</sup>, Miska M. Hannuksela<sup>b</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

## ABSTRACT

Downsampling applied to texture views prior to encoding can increase the subjective quality of decoded video. In our study, we show that spatial resolution selection based on traditional pixel-based distortion metrics, such as Mean Square Error (MSE) is weakly correlated with the resolution selection based on subjective quality of coded video. To overcome this problem, we propose a novel frequency-based distortion metric which is shown to resemble subjective quality of coded video more accurately compared to conventionally used MSE-based metric.

**Index Terms**— MVC, resolution adjustment, objective quality metrics, subjective assessment, frequency power spectrum

## 1. INTRODUCTION

3D video coding standardization is a recent activity targeting at enabling a variety of display types, including autostereoscopic multiview displays and stereoscopic displays, as well as user-adjustable depth perception. To enable this functionality, multiple high-quality views shall be available in decoder/display side. Due to the natural limitations of content production and content distribution technologies, capturing and delivery of a large number of high-quality views to user side is a very challenging task under the current video coding technologies. To assess available solutions for this challenge, the Moving Picture Experts Group (MPEG) issued a Call for Proposals for 3D video coding technologies (hereafter referred to as the 3DV CfP) [1] which enables rendering of a selectable number of views within a certain viewing range without increasing the required bitrate compared to conventional bandwidth. More than 20 solutions were submitted to the CfP and they were evaluated through a rigorous formal subjective quality assessment performed by the MPEG and its partners.

A significant number of responses to the CfP utilized the Multiview Video plus Depth (MVD) data format and were based on the H.264/MVC video coding standard [2]. The MVD data format consists of natural texture image and its associated depth map data image. The use of MVD data format and Depth-Image-Based Rendering (DIBR) algorithms [3] at the decoder side allows rendering required

number of intermediate views from limited input views. However these views (both texture and depth) should be encoded and transmitted to the decoder.

The H.264/MVC is the state-of-the-art coding standard in the field of multiview video coding (MVC) which utilizes inter-view and temporal redundancies in multiview data. However, the resulting bitrate of MVC coded MVD data (texture and depth views) exceeds the bandwidth reserved for conventional 2D video services. As a result, significant research was done to further decrease the bitrate while preserving subjective quality of decoded views and preserving the compatibility with existing H.264/MVC video coding technology.

Adaptive spatial resolution adjustment for coded video data is a potential approach to decrease the bitrate. If the same encoding parameters are utilized, downsampling of video data prior to encoding leads to bitrate reduction. In this design, the overall system distortion is a combination of conventional coding distortion and reduction of high frequency components due to low pass filtering introduced by downsampling. The video coding system should be designed properly to balance these distortions in order to achieve a subjectively superior visual quality of decoded video data.

The spatial downsampling proposed in [4, 5] improves compression at low bitrates. An adaptive decision is made for appropriate downsampling and quantization mode according to local visual significance. The downsampling ratio is automatically adjusted from 1/4 to 1 according to local image contents. Authors in [6] proposed an adaptive downscaling ratio decision approach for better compression of multiview video. The proposed method is based on a trade-off between the distortion introduced by downsampling and distortion introduced by quantization. The results indicated that using bit-rate adaptive mixed spatial resolution coding for both views and depth maps can achieve savings in bit-rate, compared to Full Resolution (FR) multiview coding when the quality of synthesized views is considered. In [7] authors utilized adaptive downsampling to improve performance of H.246/AVC video coding. In this work, it is proposed to optimize the spatial resolution through a rate distortion optimization, where distortion of downsampling and coding processes were averaged.

In this paper, we perform a set of subjective tests showing that MSE-based resolution selection cannot

estimate the subjective results accurately. Hence, a novel algorithm for adaptive spatial resolution selection based on frequency-based distortion metric is presented. Results prove that this method is capable of better estimating the subjective quality comparing MSE-based approach.

The rest of paper is organized as follows. Section 2 describes proposed methods. The test material, setup, and results are presented in sections 3 and 4 for objective and subjective experiments, respectively. Section 5 discusses the results. Finally, conclusions are given in Section 6.

## 2. PROPOSED SPATIAL RESOLUTION SELECTION METHODS

The level of distortions introduced by lossy video coding systems is typically controlled by a Quantization Parameter (QP) where higher QP corresponds with low bitrates but higher coding distortions. In the case of multi-resolution encoding and under constrained bitrate, different QP values are associated with selected resolutions. This association between different resolutions and QPs, providing a same target bitrate, can be estimated in advance and specified to the encoder through a properly designed lookup table. In such design, video data at lower spatial resolution can be coded at a lower QP under the same bitrate constrain and less coding distortions are introduced to coded data. However, process of resolution rescaling introduces its own distortion through a low pass filtering of input data. To solve this rate-distortion optimization problem, encoder should take both of these distortions in consideration. In this paper we present two methods for encoder to make decision on the spatial resolution of texture data under constrained bitrate:

- 1) Mean Square Error based method
- 2) Frequency Power Spectrum based method

The two proposed methods are described in detail in following sub-sections.

### 2.1 Pixel-based distortion metric

In this method the MSE over FR encoded image is calculated against the original image. For downsampled schemes, the encoded image with different downsampling ratios is upsampled to FR and then the MSE is calculated against the original. Considering that under the same bitrate constrain, different resolutions provide different MSE values, therefore, the resolution providing the least MSE value will be selected as the candidate which should be utilized for encoding. In this step, we consider a fine interval for MSE values in which a lower resolution is preferred. In other words, if the MSE value of the lower resolution is in a predefined and fixed interval of MSE values of a higher resolution, the lower resolution will be selected. Selecting a lower resolution favors a lower computational complexity.

### 2.2 Frequency-based distortion metric

Our approach is based on the assumption that image quality degradation caused by downsampling and coding can be better evaluated in frequency domain, rather than in the pixel domain. Since downsampling and block-based coding with scalar quantization are both suppressing high frequencies, we can evaluate introduced degradation through analysis of high frequency components of 2D Discrete Cosine Transform (DCT) spectrum.

Let us introduce the following notation:  $F\{\}$  as a separable 2D forward DCT and  $w=(w_1, w_2)$  as coordinates of DCT coefficients. 2D DCT being performed over the whole image  $s$  size of  $M \times N$  results in 2D DCT spectrum of the same size  $M \times N$ :

$$S(w_1, w_2) = F\{s(x, y)\}, \quad (1)$$

where  $F$  is the function performing the 2D DCT transfer while  $x=0, \dots, M-1$  and  $y=0, \dots, N-1$  are spatial coordinates of the image  $s$ , and  $w_1=0, \dots, M-1$  and  $w_2=0, \dots, N-1$  are coordinates in the 2D DCT spectrum  $S$ .

Transform coefficient which are located in the right-bottom section of spectral image are associated with high frequency components (**HFC**) of image  $I$  and we select these information for further analysis as follows:

$$\begin{aligned} \mathbf{HFC}(s) &= S(w_1, w_2) \\ w_1 &= T1 \dots M-1, \\ w_2 &= T2 \dots N-1 \end{aligned} \quad (2)$$

where terms  $T1$  and  $T2$  are boundaries that specify **HFC** in horizontal and vertical directions of 2D DCT spectrum, respectively.

In our method we compare **HFC** of 2D DCT coefficients computed for the following image:

- $UF$ : Uncompressed image at the Full Resolution
- $CF$ : Compressed image at the Full Resolution
- $CL$ : Compressed image at Low Resolution

Note that original image is downsampled prior to encoding and upsampled to FR after decoding to produce  $CL$ .

Each of these images undergo 2D DCT and **HFC** coefficients for  $CF$  and  $CL$  spectral images are compared against the **HFC** of the  $UF$  image:

$$\begin{aligned} dCF(w_1, w_2) &= \mathbf{HFC}(UF(w_1, w_2)) - \mathbf{HFC}(CF(w_1, w_2)) \\ dCL(w_1, w_2) &= \mathbf{HFC}(UF(w_1, w_2)) - \mathbf{HFC}(CL(w_1, w_2)) \end{aligned} \quad (3)$$

The differential spectral images  $dCF$  and  $dCL$  are computed coefficient-wise for all transform coefficients that belong to the specified **HFC**. Since transform coefficients of  $dCF$  and  $dCL$  are computed over the entire image  $s$ , a large number of them would have magnitude close to zero. These coefficients would not reflect noticeable components of the image  $s$ , but their cumulative magnitude might affect the decision making. In order to avoid this, we filter  $dCF$  and

$dCL$  with commonly used in transform-based filtering hard-thresholding [8]. This non-linear filtering operation  $T\{\}$  is performed over each transform coefficient of  $dCF$  and  $dCL$  as following:

$$T\{Y(w)\} = \begin{cases} Y(w), & |Y(w)| \geq T3 \\ 0, & \text{else} \end{cases}, \quad (4)$$

where  $Y(w)$  is original transform coefficient, and  $T(Y(w))$  filtered transform coefficient and  $T3$  is a threshold specifying an expected level of the noise present in the current image.

Following the filtering, we compute arithmetic mean of transform coefficients presented in  $T(dCF)$  and  $T(dCL)$  and utilize this value as a distortion metric.

$$\text{cost}(CF) = \frac{1}{n} \cdot \sum_{w_2=T_2}^{N-1} \sum_{w_1=T_1}^{M-1} T(dCF(w_1, w_2)) \quad (5)$$

$$\text{cost}(CL) = \frac{1}{n} \cdot \sum_{w_2=T_2}^{N-1} \sum_{w_1=T_1}^{M-1} T(dCL(w_1, w_2)) \quad (6)$$

where term  $n$  determines number of samples within **HFC** and computed as:  $n = (N - T_2 - 1) \cdot (M - T_1 - 1)$

Optimal resolution for coded image is selected as such that provide minimal cost of the metric presented in (5):

$$\text{resolution} = \arg \min_{\text{cost}} (\text{cost}(CF), \text{cost}(CL)) \quad (7)$$

### 3. OBJECTIVE EXPERIMENTS

#### 3.1 Test material and setup

Test sequences and input views utilized in this study are the same as specified in 3DV CfP for case C2 [1]. Modified JM 17.2 reference software [10] with extended multiview profile was utilized for encoding multiview texture data. Four Rate Points (RP) specified in the CfP were utilized for the encoding procedure.

The content of the sequences remains relatively similar, hence, only the statistics of the first frame are utilized in this study. However, this method can be easily extended to be utilized at Group of Picture (GOP) levels or scene cuts. If the codec supports the change on spatial resolution of frames through the encoding process, it might be subjectively beneficial to utilize the proposed method in scene cuts. Utilization of first frame statistics is due to similar content of each sequence and controlling the increase of complexity.

Considering that constant QP settings were required in the CfP, a target bitrate was met by coding FR scheme with different QP values and choosing the QP value that produced the closest bitrate to the bitrate point given in the CfP. Under the same bitrate constraint, downsampling with lower resolutions enables encoding with lower QP values

TABLE 1. SPATIAL RESOLUTION SELECTION BASED ON MSE-BASED METHOD

	Rate Points			
	RP1	RP2	RP3	RP4
Poznan Hall2	FR	FR	FR	FR
Poznan Street	FR	FR	FR	FR
Undo Dancer	FR	FR	FR	FR
GT_Fly	FR	FR	FR	FR
Kendo	1/2	1/2	3/4	3/4
Balloons	1/2	1/2	3/4	3/4
Lovebird1	1/2	3/4	3/4	3/4
Newspaper	1/2	1/2	3/4	3/4

compared to the QP utilized in FR encoding. Based on our statistical results, the ratios between QP values for downsampling ratios of 3/4 and 1/2 are 0.88 and 0.74, respectively. QP values around this estimated value for lower resolutions were tested to achieve the closest bitrate to the given bitrate point in the CfP.

#### 3.2 Results of MSE-based method

The MSE method resulted in Rate Distortion (RD) curves presenting the distortion by MSE. The lowest MSE per specific bitrate and encoding scheme is selected considering an interval equal to 5% as presented in subsection 2.1. Resolution selection based on MSE RD curves is presented in Table 1 where 1/2, 3/4, and FR present schemes where the sequences have resolution of 1/2, 3/4, and FR, respectively.

Table 1 shows that the MSE-based method resulted in the selection of 1/2 or 3/4 resolution for the 1024×768 sequences, while FR was consistently selected for the 1920×1088 sequences. In a subjective assessment of expert viewers, a resolution lower than FR was generally preferred not only for the 1024×768 sequences but also for the 1920×1088 sequences, when the viewing conditions of the CfP were used. This finding was also supported by the results of the CfP [11] as follows. We submitted coded sequences using the resolutions in Table 1. The same codec was used to encode sequences of different resolutions; hence the compression performance should be approximately equivalent regardless of the resolution. We compared the subjective evaluation results of our submission to the H.264/MVC anchor bitstreams by linearly interpolating the bitrates where H.264/MVC anchor results gave the same subjective quality as our submission in RP1 and RP2. It was found that the average bitrate reduction of RP1 and RP2 yielding the same subjective quality was about 20 percent units higher for the 1024×768 sequences in the C2 coding conditions. Comparing this bitrate reduction to that for the 1920×1088 sequences gave indications that an appropriate spatial resolution selection played an essential role in the subjective quality of the 1024×768 sequences and that the subjective quality of coded 1920×1088 sequences could be improved by downsampling.



TABLE 2. SPATIAL RESOLUTION SELECTION BASED FREQUENCY-BASED APPROACH

	Rate Points			
	RP1	RP2	RP3	RP4
Poznan Hall2	1/2	1/2	1/2	1/2
Poznan Street	1/2	1/2	1/2	1/2
Undo Dancer	1/2	1/2	1/2	1/2
GT_Fly	1/2	1/2	1/2	1/2
Kendo	1/2	1/2	1/2	1/2
Balloons	1/2	1/2	1/2	1/2
Lovebird1	1/2	FR	FR	FR
Newspaper	1/2	1/2	1/2	1/2

### 3.3 Results of the frequency-based method

Resolution selection based on the distortion metric presented in sub-section 2.2 is reported in Table 2. The thresholds we used in our experiment are  $T1 = 0.65 * width$ ,  $T2 = 0.65 * height$ , and  $T3 = 0.2 * HFC(UF)$  but the scheme is quite flexible to these thresholds. Note that these results differ from those achieved by MSE-based method (see Table 1).

Results in Table 2 show that the proposed metric favored selection of the 1/2 resolution consistently for the 1920x1088 sequences. As explained in the previous sub-section, such selection of resolutions was supported by expert viewing and also the subjective evaluation results of the CfP suggested that a lower resolution than FR could be appropriate for the 1920x1088 sequences. Nevertheless, we wanted to verify the resolutions provided by the proposed method through a systematic subjective test as explained in Section 4. Frequency based distortion metric-based method failed to select the resolution with the highest subjective quality for Lovebird1. It might be due to relatively higher (~2.5 times) cost(CL) value compared to the rest of 1024x768 sequences. The higher cost(CL) might be because of false edges due to the original sequence having JPEG-like blocking artifacts. This means downsampling eliminated more high frequency components for Lovebird1.

## 4. SUBJECTIVE EXPERIMENT

Subjective assessment was performed on three out of four 1920x1088 sequences. The input views and synthesized views utilized in our experiment are the same as specified in 3DV CfP for case C2 [1].

The same encoder as introduced in sub-section 3.1 was utilized for encoding multiview texture data and the following coding scenarios were evaluated:

- Full Resolution Scheme (FRS): 3DV coding on full resolution input
- Downsampled Scheme 1 (DS1): 3DV coding on downsampled texture with downsampling ratio 3/4 applied to both directions

- Downsampled Scheme 2 (DS2): 3DV coding on downsampled texture with downsampling ratio 1/2 applied to both directions

Each of these schemes produced a bit stream associated with rate points RP3 and RP1 given in 3DV CfP.

### 4.1 Test Procedure and Participants

Subjective viewing was conducted according to the 3DV CfP specification [1]. The 46'' Hyundai stereo display with passive glasses was utilized for displaying of the test material. The viewing distance was equal to 4 times the displayed image height (2.29m for HD sequences).

Subjective quality assessment was done according to Double Stimulus Impairment Scale (DSIS) method [12] with discrete unlabeled quality scale from 0 to 10 was used for quality assessment. Prior to the actual test, subjects were familiarized with test task, test sequences and with the variation in quality they could expect in the actual tests. The viewers were instructed that 0 stands for the lowest quality and 10 for the highest.

Prior to the participation in subjective viewing experiment, candidates were subject to a thorough vision screening. Candidates who did not pass the criteria (near and far vision, Landolt chart) of 20/40 visual acuity with each eye or color vision (Ishihara) were rejected. All participants had a stereoscopic acuity of at least 60 arc sec.

Subjective viewing was conducted with 30 subjects, (19 female, 11 male), aged between 18-29 years (mean: 23.7). The majority of the candidates (90%) were considered naïve as they did not work or study in fields related to information technology or video processing.

### 4.2 Subjective results

Subjective test results are depicted in Figure 1. It can be judged from the mean scores and confidence intervals presented in Figure 1 that subjective quality of DS2, associated to the lowest resolution, tends to be higher

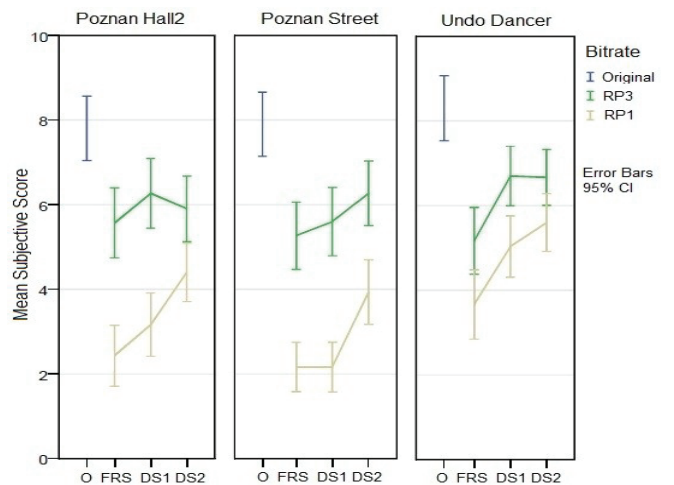


Figure 1. Subjective results for different encoding schemes

TABLE 3. SPATIAL RESOLUTION SELECTION BASED ON STATISTICAL SIGNIFICANCE ANALYSIS ON SUBJECTIVE RESULTS

	Rate Points	
	RP1	RP3
Poznan Hall2	1/2	1/2, 3/4, FR
Poznan Street	1/2	1/2
Undo Dancer	1/2	1/2, 3/4

compared to other schemes. The observation on significant differences between the encoding schemes was further analyzed using statistical analysis as presented in the paragraphs below.

Non-parametric statistical analysis methods, Friedman's and Wilcoxon's tests, were used as the data did not reach normal distribution (Kolmogorov-Smirnov:  $p < .05$ ). Friedman's test is applicable to measure differences between several and Wilcoxon's test between two related and ordinal data sets [13]. A significance level of  $p < .05$  was used.

The following conclusions were obtained with statistical significance analysis presented above. In lower bitrates, DS2 has always better subjective results. In higher bitrates, all schemes have a similar performance for Poznan Hall2. In Poznan Street, DS2 has significantly a better subjective quality while DS1 and DS2 have a similar subjective quality for Undo Dancer and both are performing better than FRS. These results are reported in Table 3.

## 5. DISCUSSION

In this section, the objective results of the proposed methods are compared with subjective results available on a sub-set of test material. The subjective results are used as a reference and the performance of MSE- and DCT-based methods is evaluated based on similarity of their results with subjective results i.e. the more accurately estimating the subjective results, the better performing the method.

First, we compared the objective results achieved by MSE method with subjective results on available subset of test material. We noticed that MSE is not an appropriate metric to predict the subjective quality since only one of the resolution selections made by this method were aligned with subjective results. MSE results in all cases for HD sequences were favored to select the encoding schemes with FR while subjective results showed otherwise.

Second, resolution selection achieved by proposed method was compared with selection based on subjective test. In all cases the proposed method succeeded to estimate the subjective results correctly.

## 6. CONCLUSIONS

This paper tackled the problem of adaptive spatial resolution selection by comparing two methods. First, MSE value was calculated for FR and lower resolutions. The

resolution with the smallest average MSE value was selected as the candidate to have the best subjective quality. This selection was compared then with subjective results on a subset of test material, revealing that the MSE-based method is not able to estimate the subjective quality accurately (one out of the six cases were estimated correctly). To solve this problem an objective metric based on FPS was described. The results confirmed that utilization of this algorithm succeeded to select the resolution with the best subjective quality whenever the subjective quality assessment results were available (all cases were estimated correctly). Hence, the proposed method is a potential candidate metric to select the resolution of the texture view prior to encoding by which the best perceived quality is assured.

## ACKNOWLEDGMENT

The authors would like to thank Timo Utriainen, Emilia Pesonen, and Satu Jumisko-Pyykkö from the laboratory of the Human-Centered Technology of Tampere University of Technology for performing and providing the subjective results. Moreover, the authors thank Prof. M. Domański, et al. for providing Poznan sequences and their camera parameters [9].

## REFERENCES

- [1] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Geneva, Switzerland, March 2011.
- [2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.
- [3] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in Proc. SPIE Conf. 5291, CA, U.S.A., Jan. 2004, pp. 93–104.
- [4] W. Lin and D. Li, "Adaptive downsampling to improve image compression at low bit rates," IEEE Trans. Image Process., vol. 15, no. 9, pp. 2513–2521, Sep. 2006.
- [5] V.A. Nguyen, Y.P. Tan, W.S. Lin, "Adaptive downsampling/upsampling for better video compression at low bit rate," IEEE ISCAS, pp.1624-1647, May 2008.
- [6] E. Ekmekcioglu, S. T. Worrall, and A. M. Kondoz, "Bit-rate adaptive downsampling for the coding of multiview video with depth information," in Proc. 3DTV Conf., Istanbul, Turkey, May 2008, pp. 137–140.
- [7] Ren-Jie Wang, Ming-Chen Chien, and Pao-Chi Chang, "Adaptive downsampling video coding," Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7542, 2010
- [8] R. Oktem, L. Yaroslavsky and K. Egiazarian, "Signal and Image Denoising in Transform Domain and Wavelet Shrinkage: A Comparative Study", in Proc. of EUSIPCO'98, Sept. 1998.
- [9] M. Domański, et al, "Poznan Multiview Video Test Sequences and Camera Parameters", MPEG 2009/M17050, October, 2009.
- [10] JM reference software: <http://iphome.hhi.de/suehring/tml/download>
- [11] [http://mpeg.chiariglione.org/working\\_documents/explorations/3dav/3d-test-report.zip](http://mpeg.chiariglione.org/working_documents/explorations/3dav/3d-test-report.zip)
- [12] ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002.
- [13] H. Cooligan "Research methods and statistics in psychology," (4th ed.). London: Arrowsmith., 2004.

- [P8] **P. Aflaki**, D. Rusanovskyy, T. Utriainen, E. Pesonen, M. M. Hannuksela, S. Jumisko-Pyykk, and M. Gabbouj; “Study of asymmetric quality between coded views in depth-enhanced multiview video coding, ” International Conference on 3D Imaging (IC3D), Liege, Belgium, December. 2011.

© IEEE, 2011, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# STUDY OF ASYMMETRIC QUALITY BETWEEN CODED VIEWS IN DEPTH-ENHANCED MULTIVIEW VIDEO CODING

*Payman Aflaki<sup>a</sup>, Dmytro Rusanovskyy<sup>b</sup>, Timo Utriainen<sup>c</sup>, Emilia Pesonen<sup>c</sup>, Miska M. Hannuksela<sup>b</sup>, Satu Jumisko-Pyykko<sup>c</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

<sup>c</sup>Human-Centered Technology, Tampere University of Technology, Tampere, Finland;

## ABSTRACT

Depth-enhanced multiview video formats, such as the multiview video plus depth (MVD) format, enable a natural 3D visual experience which cannot be brought by traditional 2D or stereo video services. In this paper we studied an asymmetric MVD technique for coding of three views that enabled rendering of the same bitstream on stereoscopic displays and multiview autostereoscopic displays. A larger share of bitrate was allocated to a central view, whereas two side views were coded at lower quality. The three decoded views were used by a Depth-Image-Based Rendering algorithm (DIBR) to produce virtual intermediate views. A stereopair at a suitable separation for viewing on a stereoscopic display was selected among the synthesized views. A large-scale subjective assessment of the selected synthesized stereopair was performed. A bitrate reduction of 20% on average and up to 22% was achieved with no penalties on subjective perceived quality. In addition, our analysis shows that a similar bitrate reduction gain with no difference in subjective quality can be achieved in multiview autostereoscopic display scenario

**Index Terms**— 3DV, MVC, asymmetric quality multiview video, subjective assessment.

## 1. INTRODUCTION

3D video coding standardization in the Moving Picture Experts Group (MPEG) is a recent activity targeting at enabling a variety of display types and preferences including varying baseline to adjust the depth perception. Another important target of the MPEG 3DV standardization is the support for multiview autostereoscopic displays (ASDs), thus many high-quality views shall be available in decoder/display side prior to displaying. Due to the natural limitations of content production and content distribution technologies, there is no way that a large number of views can be delivered to user with existing video compression standards. Therefore, MPEG issued a Call for Proposals for 3D video coding (hereafter referred to as the 3DV CfP) [1] for a new standard which enables rendering of a selectable number of views without increasing the required bitrate.

One candidate for 3D video presentation is ASD, emitting more than one stereopairs at a time enabling glass-less 3D perception. However, the ASD technology ensures that subjects observe only one stereopair at a time and subjects can change their viewpoint and consequently observe different stereopairs of the same 3D scene. For this purpose many views should be available for the autostereoscopic display. A multiview video plus depth (MVD) format [2], where each video data pixel is associated with a corresponding depth map value, allows reducing the input data for the 3DV systems significantly, since most of the views will be rendered from the available decoded views and depth maps using a DIBR [3] algorithm. Autostereoscopic displays provide a larger viewing angle and as a result a wider camera separation is needed. Hence, as proposed by the 3DV CfP, a 3-view MVD coding scenario is suitable for creation of a wide range of required views for multiview ASD rendering while a suitable pair of synthesized views can also be used for rendering on a stereoscopic display.

The 3DV CfP, in addition to the data format, targets the development of new 3DV coding technologies. The MVD format can be considered as one of the most potential approaches for the 3DV CfP. The Multiview Video Coding extension of the Advanced Video Coding standard (H.264/MVC) [4] is the state-of-the-art standard in the field of multiview video coding. H.264/MVC can be applied for coding of MVD data, for example by coding the multiview texture video as one H.264/MVC bitstream and the respective multiview depth video as another H.264/MVC bitstream. Despite the high coding efficiency of H.264/MVC, the resulting bitrate of coded multiview data exceeds the bandwidth reserved for conventional 2D video services by a great margin. As a result, more research has been focused on possible approaches to reduce the bitrate of coded multiview video while preserving subjective quality of decoded views and preserving the compatibility with the H.264/MVC standard.

Asymmetric stereoscopic video has been researched as one of the possible solutions to reduce the bitrate and/or computational complexity. In asymmetric stereoscopic video, the two views have different visual quality. The usage of this technique is motivated by the binocular rivalry

theory [5], which claims that the stereoscopic vision in human visual system (HVS) fuses the images of a stereopair so that the visual perceived quality is closer to that of higher quality view. Several subjective quality evaluation studies have been conducted to research the utilization of the binocular rivalry theory in stereoscopic video. For example, a set of subjective tests comparing symmetric, quality-asymmetric stereoscopic video coding were conducted in [6]. The presented results showed that subjective quality of symmetric and asymmetric stereoscopic videos provided similar quality under the same bitrate constraint.

In this paper, we study the applicability of asymmetric coding for the three-view (C3) test scenario of the 3DV CfP. We propose a 3-view coding arrangement, where the central view can be extracted for 2D viewing, a central stereopair can be derived for viewing on stereoscopic displays, and a multitude of views can be generated through DIBR for viewing on a multiview ASD. The side views are proposed to be coded at lower quality compared to the quality of the central view, referred to as “full quality”. Consequently, the proposed method yields bitrates that are significantly lower than the bitrates of the corresponding symmetric full-quality bitstreams. A subjective assessment in a typical stereoscopic viewing environment was conducted to compare the proposed scheme with a conventional symmetric scheme. The results of the subjective evaluations confirmed that a significant decrease in bitrate (20% of bitrate reduction on average for best tested scheme) was achieved with no degradation in the subjective quality. Moreover, we objectively confirmed the applicability and efficiency of the proposed asymmetric scheme for ASD utilization.

The rest of the paper is organized as follows. Section 2 presents the utilized quality-asymmetric MVD coding scheme. The performed experiments are described in Section 3, while Section 4 provides the results. Finally, the paper concludes in Section 5.

## 2. ASYMMETRIC CODING FOR 3D MULTIVIEW VIDEO

This Section introduces the proposed three-view MVD coding method that utilizes asymmetric transform-domain quantization between views. In order to provide grounds for the proposed coding method, a review of asymmetric stereoscopic video coding and a description of rendering of 3D video on different types of displays are given in Sections 2.1 and 2.2, respectively. Then, we present the proposed coding method in Section 2.3.

### 2.1 Asymmetric Stereoscopic Video Coding

Asymmetric stereoscopic video coding includes a large variety of encoding schemes which provide a quality difference between two views. If different encodings are applied for left and right view, the coding artefacts of one method in the lower quality view can be masked by details presented in the higher quality view. It is evident that there

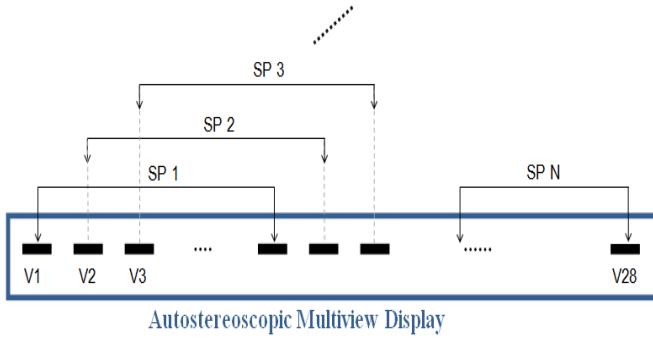
are limits on the amount of asymmetry that binocular fusion can successfully mask so that the perceived quality is closer to the quality of the higher-fidelity view. Asymmetry between the two views can be achieved by one or more of the following methods:

- a) Mixed-resolution (MR) stereoscopic video coding, first introduced in [7], also referred to as resolution-asymmetric stereoscopic video coding. One of the views is low-pass filtered and hence has a smaller amount of spatial details or a lower spatial resolution. Furthermore, the low-pass filtered view is usually sampled with a coarser sampling grid, i.e., represented by fewer pixels.
- b) Mixed-resolution chroma sampling [8]. The chroma pictures of one view are represented by fewer samples than the respective chroma pictures of the other view.
- c) Asymmetric sample-domain quantization [9]. The sample values of the two views are quantized with a different step size. For example, the luma samples of one view may be represented with the range of 0 to 255 (i.e., 8 bits per sample) while the range may be scaled to the range of 0 to 159 for the second view. Thanks to fewer quantization steps, the second view can be compressed with a higher ratio compared to the first view.
- d) Asymmetric transform-domain quantization. The transform coefficients of the two views are quantized with a different step size. As a result, one of the views has a lower fidelity and may be subject to a greater amount of visible coding artifacts, such as blocking and ringing.
- e) A combination of different encoding techniques above.

It was found in [10] that the perceived quality of video clips produced using asymmetric transform-domain quantization was approximately equal to the average of the perceived qualities of the two views individually. The impact of the quantization of transform coefficients was verified in [11], where it was concluded that the perceived quality of coded equal-resolution stereo image pairs was approximately the average of the perceived qualities of the high-quality image and the low-quality image of the stereopairs. Furthermore, the same conclusion of the perceived quality of asymmetric transform-domain quantization was also reached in [6].

### 2.2 Rendering of 3D Video on Stereoscopic and Autostereoscopic Displays

More than two views are rendered simultaneously on a multiview autostereoscopic display. As stated earlier, many multiview autostereoscopic displays provide a wider separation of views as typical stereoscopic displays. At a given time a user sees two views, but by changing the head position the user is able to look at other stereopairs of the rendered views. For example, the display chosen for the 3DV CfP, Dimenco BDL5231V3D, renders 28 views. Hence, to address the required wider separation of views



**Fig. 1.** Perceivable sliding stereopair for ASD

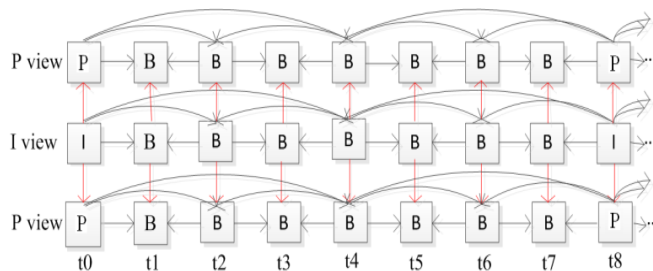
required by the Dimenco display and many other autostereoscopic displays, the 3DV CfP includes a 3-view scenario, which is suitable for creation of a wide range of required views.

Fig. 1 illustrates the fact the user sees two views (stereopair SP  $x$ ) at a time out of the 28 views of the Dimenco display and can choose his/her head position among the possible stereopairs. The difference of view numbers in a stereopair depends on several factors such as the interpupillary distance, the viewing distance, and the rendering parameters of the display. We assumed that views  $N$  and  $N+8$  out of the 28 views displayed by the Dimenco display could be considered to form a stereopair in a typical multiview autostereoscopic viewing situation.

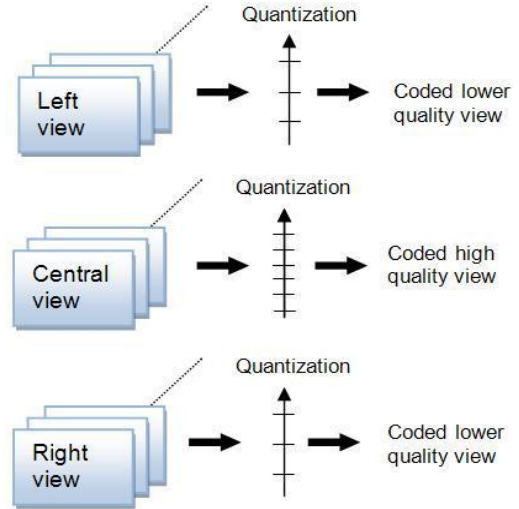
When a 3-view bitstream is adapted for comfortable viewing on a stereoscopic display, two views at a suitable separation have to be extracted. It can be assumed that content providers would appreciate that the mid-most stereopair is selected for such stereoscopic viewing, hence providing a “central” viewpoint to the content, rather than picking views from either side of the 3-view setting. Therefore the decoded views in a 3-view bitstream are not displayed as such for rendering on a stereoscopic display, but a suitable stereopair is synthesized from the decoded views.

### 2.3 Proposed Asymmetric Three-View MVD Coding

Inspired by the promising results of asymmetric stereoscopic video coding (see Section 2.1) we wanted to design an asymmetric coding scheme for 3-view MVD format. We started off with the following requirements and assumptions:



**Fig. 2.** PIP interview prediction structures







**Fig. 3.** Proposed asymmetric MVC scheme for 3DV coding

1. H.264/AVC decoders must be able to extract 2D video from the 3-view MVD bitstreams in order to obtain compatibility with existing decoders.
2. The extracted 2D video should be the central view of the content as it is likely to represent the captured 3D scene more appropriately than the side views.
3. No compromise on the 2D video quality should be made in the asymmetric coding, as for the time being 2D viewing is still more common than 3D viewing.
4. Two midmost views at a suitable separation should be generated for viewing on stereoscopic displays (see Section 2.2). The quality of the midmost views for stereoscopic viewing should not be compromised compared to the corresponding views obtained from symmetric 3-view MVD bitstreams.
5. Such number of intermediate views that suits the multiview ASD being used should be able to be created from the decoded 3-view MVD bitstream. The average quality of the perceived stereopairs on the multiview ASD should be similar to that obtained from symmetric 3-view MVD bitstreams.

In order to respond to the requirements 1 to 3, we decided to use H.264/MVC independently for texture and depth views and select the PIP inter-view prediction structure for encoding (see Fig. 2). In this structure the central view is coded with H.264/AVC and utilized as reference for coding of the two side views. The transform coefficients of the central (base) view are quantized using a fine quantization parameter ( $QP_0$ ) while the side views are quantized more coarsely with a quantization parameter  $QP_0 + \Delta QP$  (Fig. 3). This will result in a lower quality of the side views compared to the quality of the base view and consequently brings a bitrate reduction comparing the symmetric quality case where all views are encoded using the same  $QP_0$ . Alternatively, it is possible to encode an asymmetric MVD bitstream with the same bitrate as a

Table 1. Sequences and their characteristics

Screenshot	Sequence	Resolution	Frames	Framerate
	Poznan Hall2	1920x1080	250	25
	Undo Dancer	1920x1080	250	25
	Balloons	1024x768	300	30
	Newspaper	1024x768	300	30

symmetric MVD bitstream by coding the central view with a QP value lower than the QP for the symmetric MVD.

Asymmetric transform-domain quantization was chosen because it enables the realization of the proposed coding scheme with the H.264/MVC standard as an encoding method without changes to the bitstream format or the decoder. In principle, a similar asymmetric MVD coding scheme could utilize also utilize the other types of asymmetry listed in Section 2.1.

In the rest of this paper, we study how the requirements 4 and 5 above are met with the proposed asymmetric MVD coding.

### 3. DESCRIPTION OF THE EXPERIMENTS

As introduced in the Section 2.3, we performed experiments to clarify whether the proposed asymmetric three-view MVD coding scheme is beneficial for stereoscopic viewing and multiview autostereoscopic viewing when compared to symmetric MVD coding. We decided to carry out a large-scale systematic subjective evaluation experiment for the stereoscopic viewing, because impact of view synthesis on the perceived quality has not been explored earlier in a similar viewing scenario as much as we are aware and the usage of objective metrics would have been therefore questionable. With the results obtained from the stereoscopic viewing test, we were able to make assumptions on the behavior of objective metrics with respect to perceived quality and hence generalize the findings of the stereoscopic viewing assessment for multiview autostereoscopic displays.

In Section 3.1 we introduce the test material, the compared coding scenarios, and the test stimuli preparation from the decoded bitstreams. The setup and procedure of the

Table 2. Test sequences, input views, and synthesized views

Sequence	Input views	Synthesized views
Poznan Hall2	7-6-5	6.25-5.75
Undo Dancer	1-5-9	4-6
Balloons	1-3-5	2.5-3.5
Newspaper	2-4-6	3.5-4.5

Table 3. Tested rate points an corresponding QP settings (SS)

Sequence	R2		R4	
	QP0	Bitrate, Kbps	QP0	Bitrate, Kbps
Poznan Hall2	<b>36</b>	715.3	<b>28</b>	1747.1
Undo Dancer	<b>40</b>	1369.4	<b>36</b>	2296.8
Balloons	<b>44</b>	374.8	<b>35</b>	973.9
Newspaper	<b>40</b>	553.7	<b>34</b>	1049.3

subjective viewing experience evaluation on a stereoscopic display are presented in Section 3.2.

#### 3.1 Test stimuli

In this section the detailed steps of test material preparation is described. Four sequences, Undo Dancer, Newspaper, Poznan Hall2 [12], and Balloons, included in the 3DV CfP test set were used. Basic parameters of these sequences, such as resolution, and frame rates are given in Table 1.

The selected input and output (synthesized) views in our experiments are shown in Table 2. Note that input views for all tested sequences are the same as specified in the 3DV CfP for the C3 case. Based on the discussion presented in sub-section 2.3, we chose a stereopair for each sequence from the center of the three coded views such that the baseline difference of the chosen views suits viewing on a stereoscopic display. It is noted that the view separation of the stereopairs specified in the 3DV CfP for C3 is approximately half of a conventional stereo baseline. Such narrow view separation was chosen to mimic the viewing conditions on typical autostereoscopic multiview displays. As we targeted for a different use case, the synthesized views utilized in stereo viewing were selected differently from the MPEG 3DV CfP.

The proposed asymmetric MVC scheme was implemented on the top of JM 17.2 reference software [13]. The software was configured to produce test materials with three MVC schemes:

- Symmetric Scheme (SS): all views were coded with equal  $QP = QP_0$ .
- Asymmetric Scheme 1 (AS1), the central view was coded at  $QP = QP_0$ , and the side views were coded with  $QP = QP_0 + 2$
- Asymmetric Scheme 2 (AS2): the central view was coded at  $QP = QP_0$ , and the side views were coded with  $QP = QP_0 + 4$

The SS bitstreams were encoded for two bitrate points referred to as R2 and R4 where the QP values were selected equal to those of the H.264/AVC anchor encoding of the 3DV CfP. The bitrates and the utilized QP values of these SS bitstreams are provided in Table 3. The other encoding settings were chosen to comply with the requirements of the 3DV CfP.

The tested schemes (SS, AS1 and AS2) produced identical bitrates for the coded central views, whereas the bitrates for the coded side views in AS1 and AS2 were significantly reduced compared to the SS. An average total

Table 4. Bitrate reduction of proposed asymmetric method compared to symmetric coding method

Sequence	AS1 (%)	AS2 (%)
Poznan Hall2	12.9	22.1
Undo Dancer	10.1	16.8
Balloons	12.4	19.6
Newspaper	13.1	22.2
<b>Average</b>	<b>12.6</b>	<b>20.2</b>

bitrate reduction compared to SS was 12.6% for AS1 and 20.2% for AS2. The detailed bitrate reduction in a sequence basis is reported in Table 4.

Stereopairs, as specified in Table 2, were synthesized for each of tested schemes (SS, AS1, AS2). The view synthesis was performed with VSRS software, version 3.5 [14]. We utilized VSRS configuration files and camera parameters information, as they are specified in the MPEG 3DV CfP.

For the multiview autostereoscopic viewing we assumed the use of Dimenco BDL5231V3D autostereoscopic display as specified in the 3DV CfP. The 28 views required by the Dimenco display were produced as follows. First, we synthesized the views as described in the 3DV CfP, resulting into 49 and 33 coded or synthesized views for Newspaper and the rest of the sequences, respectively, and then we picked the 28 mid-most ones for rendering.

### 3.2 Subjective Quality Evaluation on Stereoscopic Display

25 subjects, (18 female, 7 male), aged between 19-29 years (mean: 23.9) participated in this experiment. A majority (84%) of them were considered naïve as they did not work or study in fields related to information technology, television or video processing.

The test session comprised three parts: 1) pre-test sensorial screening and demo-/psychographic data-collection, 2) actual voting using quantitative data-collection, and 3) post-test interview with qualitative data-collection. The candidates were subject to thorough vision screening. Candidates who did not pass the criterion of 20/40 (near and far vision, Landolt chart) visual acuity with each eye or color vision (Ishihara) were rejected. All participants had a stereoscopic acuity of 60 arc sec at the minimum.

The laboratory conditions were organized according to [15]. Hyundai 46-inch stereoscopic monitor model S465D with passive polarizing glasses was used, as suggested by the 3DV CfP. Furthermore, the viewing distance was four times the height of the image (2.29m for 1920×1080 and 1.63m for 1024×768 video sequences), as specified by the 3DV CfP.

The subjective test started with a combination of anchoring and training. The extremes of the quality range of the stimuli were shown to familiarize the participants with the test task, the test sequences, and the variation in quality

they could expect in the actual tests that followed. The test clips were presented one at a time in random order and appeared twice in the test session following the ITU recommendation Double Stimulus Impairment Scale (DSIS) method [15]. A discrete unlabeled quality scale from 0 to 10 was used for the rating scale. The viewers were instructed that 0 stands for the lowest quality and 10 for the highest.

The post-test sessions contained a semi-structured interview that gathered the participant’s impressions, experiences and descriptions of the visual quality to deepen the understanding behind the decisions of the participant [16]. The interview was constructed of main and supporting questions. The main question “*What kind of factors did you pay attention to while evaluating quality?*” was asked several times with slight variations during the interview. The moderator only used the terms introduced by the participant when asking the supporting questions to further clarify the answers to the main question: “*Please could you clarify if X was among the positive/negative factors or pleasant/unpleasant?*” and “*Which of the factors you mentioned was the most pleasant/unpleasant?*”.

The participant filled the Simulator Sickness Questionnaire (SSQ) [17] before and after the actual test. The questionnaire measures sickness symptoms using a weighted average of nausea, oculomotor and disorientation scores, while also calculating a composite score.

## 4. RESULTS

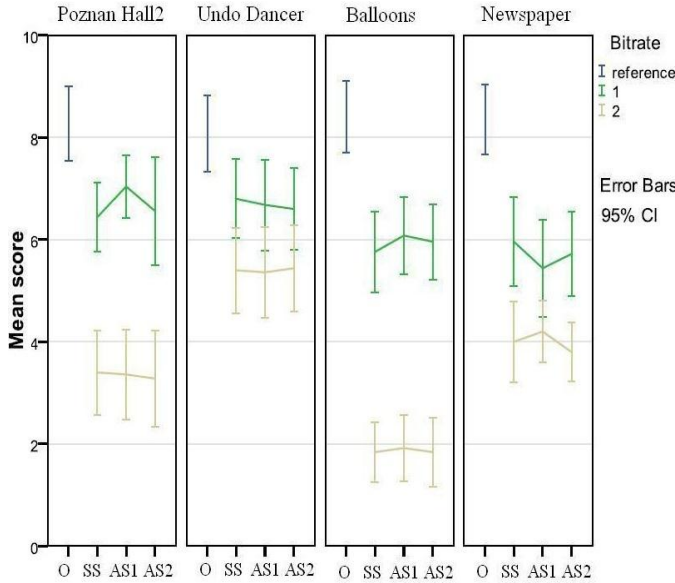
In this section we present the results and the analysis of the performed experiments. Section 4.1 presents the statistics of the quantitative viewing experience ratings on the stereoscopic display. The qualitative results obtained from the post-test interviews are summarized in Section 4.2. In Section 4.3 we analyze the obtained results through selected objective metrics and draw conclusions on the expected performance of the proposed method for multiview autostereoscopic rendering.

### 4.1 Quantitative quality evaluation

Fig. 4 shows the average and the 95% confidence interval (CI) of the subjective viewing experience ratings for all sequences in two different bitrates. The naming introduced in sub-section 3.2 is used and O stands for the original uncompressed sequences.

As can be judged from the average ratings and confidence intervals presented in Fig. 4, no significant differences were perceived between the encoding schemes for the same value of QP0. In other words, it can be observed from Fig. 4 that the described asymmetric scheme provided the same subjective quality as the symmetric scheme with the test material. However, the utilization of asymmetric coding is preferred since on average the bitrate was reduced by 12.6% and 20.2% for AS1 and AS2, compared to SS, respectively (see Table 4). The observation that there were no significant differences between the





**Fig. 4.** Viewing experience ratings (O = original uncompressed sequence, SS = symmetric coded sequence, AS1, AS2 = asymmetric transform-domain quantization between coded views)

encoding schemes was further verified using statistical analysis as presented in the paragraphs below.

Non-parametric statistical analysis methods, Friedman’s and Wilcoxon’s tests, were used as the data did not reach normal distribution (Kolmogorov-Smirnov:  $p < .05$ ). Friedman’s test is applicable to measure differences between several and Wilcoxon’s test between two related and ordinal data sets [18]. A significance level of  $p < .05$  was used unless otherwise stated below.

The results of the Friedman’s test verified that there were no significant differences between the encoding schemes in the test stimuli except that AS1 was rated higher than SS in bitrate 1 of Poznan Hall2 ( $p < .05$ ). Furthermore, no significant difference ( $p < 0.001$ ) was observed for the subjective rating of different schemes for each sequence and for both proposed bitrates in the performed Wilcoxon pairwise comparisons either.

As explained earlier, the side views in the AS1 and AS2 coding scenarios had lower objective quality as the side views in the SS coding scenario. The objective quality as measured by the average luma Peak-Signal-to-Noise Ratio (PSNR) metric for the tested coding scenarios and sequences is reported in Table 5. The applied view synthesis algorithm utilized the decoded texture and depth views that are adjacent to the view being synthesized, hence the objective quality of the synthesized stereopair of AS1 and AS2 was lower than that of SS. This objective quality difference was analyzed by first deriving the average luma PSNR of each synthesized view of ratepoint R4 against the views synthesized from uncompressed data, so called Reference View Synthesis (RVS). Then, the the average of

the PSNR values of the two views of the selected stereopair (denoted aPSNR) was taken and finally the difference (dPSNR) of aPSNR(ASx) to aPSNR(SS) was computed. The values of dPSNR are reported in Table 6. It can be seen that the dPSNR results for the synthesized stereopair of Dancer, Balloons, and Newspaper sequences contradicted with the obtained results of the subjective quality evaluation experience, hence giving an indication that two stereopairs having an average luma PSNR difference smaller than or similar to the ones reported in Table 6 may have an equal quality subjectively.

#### 4.2 The results of the post-test interview

The analysis was based on grounded theory and its wide applications to visual and audiovisual quality [16, 19]. All recorded interviews were transcribed to the text as a pre-processing step of analysis. 30% of interviews (7 participants) were used as a base for open coding (read through, extraction of meaningful sentences and coding for creating the concepts and their properties). All concepts were organized into sub-categories and they were further organized under main categories. This phase was conducted by one researcher and reviewed by another researcher. The categorization created was used in the coding of the whole data. One mention per category per person was counted and frequencies in each category were determined by counting the number of participants that described the category.

The descriptive quality of experience was composed of five main components: 1) 3D quality, 2) spatial quality, 3) temporal quality, 4) viewing task and 5) content and quality variation. The most commonly mentioned negative quality factors were associated to spatial quality (inaccuracy in general, or inaccuracy of outlines of objects and details), visibility of impairments with detectable structure, impairments during motion, and hardness or unpleasurable viewing were mentioned by more than 48% of participants. In contrast, the most commonly described positive quality of experience factors were excellence of depth impression and fluency of motion (more than 48% of participants).

Table 5. PSNR of coded views

Sequence	View	dQP0		dQP2		dQP4	
		R2	R4	R2	R4	R2	R4
Hall2	3	38.6	41.1	38.1	40.7	37.4	40.3
	4	38.6	41.1	38.6	41.1	38.6	41.1
	5	38.6	41.0	38.0	40.6	37.3	40.2
Dancer	1	30.4	32.2	30.1	31.8	29.7	31.4
	5	30.6	32.4	30.6	32.4	30.6	32.4
	9	30.6	32.4	30.2	32.0	29.8	31.5
Balloons	1	31.1	36.8	30.3	35.9	29.8	34.9
	3	31.5	37.3	31.5	37.3	31.5	37.3
	5	31.1	36.7	30.2	35.7	29.7	34.8
Newspaper	2	32.6	35.8	31.6	34.9	30.7	33.9
	4	33.1	36.2	33.1	36.2	33.1	36.2
	6	32.4	35.4	31.8	34.7	31.1	34.1

### 4.3 Objective analysis of results for ASD utilization

The results presented in Section 4.1 indicated that synthesized stereoscopic videos having a relatively small average luma PSNR difference between them appeared to have subjectively equal quality. In this section we compare the performance of different coding schemes (SS, AS1, AS2) using PSNR and show the efficiency of the proposed asymmetric MVD coding scheme for multiview ASD utilization objectively.

The PSNR analysis similar to that for the stereoscopic case was performed as follows. First, the average luma PSNR of each synthesized view of ratepoint R4 was computed against RVS. Second, the average luma PSNR values of each stereopair covering the full range of all available 28 views for the used ASD (see Fig. 1) were derived. Third, the average luma PSNR difference of the respective stereopair positions (dPSNR) between asymmetric schemes (AS1 and AS2) and SS were calculated. Table 6 provides the average dPSNR over all stereopairs visible in the Dimenco multiview ASD. It can be seen that dPSNR of the stereopair tested subjectively on the stereoscopic display is close to the mean dPSNR of the stereopairs perceivable on the ASD. This confirms that there is not likely to be a considerable difference between quality perception of 3D video in the multiview ASD presentation for the tested SS, AS1, and AS2 sequences. However, the utilization of the proposed asymmetric scheme is preferred, since a bitrate reduction up to 22% was achieved with the tested sequences. Moreover, it could be assumed that viewers tend to choose a position within a viewing cone that is more likely in the center of the cone than on either far side, because a center position allows more flexibility in head movement. Consequently, as the proposed asymmetric coding scheme favors the middle views coded or synthesized for the ASD, it could be even more favorably perceived than what an average quality measure over all possible stereopairs would indicate. Hence, we can assume that the conclusions for the stereoscopic case would hold for autostereoscopic rendering too.

We further investigated whether the PSNR calculation against RVS is a valid metric for our study. For this purpose, we considered the Undo dancer sequence for which

Table 6. dPSNR of stereopairs, AS1 and AS2 compared to SS.

“ASD, mean” = the mean dPSNR over all stereopairs of the used multiview ASD display.

“SD” = the selected stereopair for the stereoscopic display.

	AS1-dPSNR (dB)		AS2-dPSNR (dB)	
	ASD, mean	SD	ASD, mean	SD
Poznan Hall2	0.12	0.00	0.24	0.01
Dancer	0.15	0.11	0.29	0.21
Balloons	0.37	0.28	0.74	0.57
Newspaper	0.20	0.11	0.40	0.22

Table 7. Report on dPSNR against RVS and Original for available middle views

	AS1-dPSNR (dB)		AS2-dPSNR (dB)	
	ASD, mean	SD	ASD, mean	SD
Against Original	0.11	0.09	0.21	0.16
Against RVS	0.14	0.11	0.27	0.21

we had six original views (out of the total 28 displayed ones) available. The same procedure as described above was performed to calculate the dPSNR values for the available stereopairs against RVS and original views. Results, reported in Table 7, show that there is no remarkable difference between dPSNR against RVS and original views. This confirms the validity of our conclusions utilizing PSNR calculations against RVS.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper we studied a quality-asymmetric multiview-video-plus-depth coding scheme for the 3-view test scenario specified in MPEG 3DV CfP. The asymmetric quality coding was implemented through coarser transform-domain quantization for the side views, whereas the central view was coded at high quality. Decoded three views were used by a Depth Image Based Rendering algorithm to produce virtual intermediate views that enabled viewing either on stereoscopic displays or multiview autostereoscopic displays. A large-scale subjective assessment of synthesized stereopair was performed on stereoscopic displays and the results showed that a bitrate reduction of 20%, on average, was achieved with no penalties on the perceived quality when compared to coding all the views at a symmetric quality. We also analyzed through objective quality metrics that the described asymmetric coding scheme is also likely to yield subjectively equal quality at the same bitrate reduction factor compared to coding views at symmetric quality when viewed on a multiview autostereoscopic display. As a future task, we plan to verify the conclusions for the multiview autostereoscopic displays with a systematic subjective quality evaluation study.

## 6. ACKNOWLEDGEMENT

The authors would like to thank M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner for providing Poznan Hall2 sequence and Camera Parameters [12].

## 7. REFERENCES

- [1] MPEG Video and Requirement Groups, "Call for Proposals on 3D Video Coding Technology", MPEG output document N12036, Geneva, Switzerland, March 2011
- [2] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," Proc. of IEEE International Conference on Image Processing, vol. 1, pp. 201-204, Oct. 2007.
- [3] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in Proc. SPIE Conf.

- Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, CA, U.S.A., Jan. 2004, pp. 93–104.
- [4] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.
  - [5] R. Blake, "Threshold conditions for binocular rivalry," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3(2), pp. 251-257, 2001.
  - [6] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, "Subjective Study on Compressed Asymmetric Stereoscopic Video," *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Sep. 2010.
  - [7] M.G. Perkins, "Data compression of stereopairs," *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684-696, Apr. 1992
  - [8] A. Aksay, C. Bilen, G. Bozdagi Akar, "Subjective evaluation of effects of spectral and spatial redundancy reduction on stereo images," *13th European Signal Processing Conference, EUSIPCO-2005*, Turkey, Sep. 2005.
  - [9] P. Aflaki, M. M. Hannuksela, J. Hakala, J. Häkkinen, and M. Gabbouj, "Joint adaptation of spatial resolution and sample value quantization for asymmetric stereoscopic video compression: a subjective study," *Proc. International Symposium on Image and Signal Processing and Analysis*, Sep. 2011.
  - [10] W.J. Tam, "Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV," *Joint Video Team document JVT-W094*, Apr. 2007
  - [11] P. Seuntjens, L. Meesters, and A. IJsselstein, "Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Transactions on Applied Perception*, vol. 3, no. 2, pp. 96-109, Apr. 2006
  - [12] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters", *ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050*, Xian, China, October 2009.
  - [13] JM reference software:  
<http://iphone.hhi.de/suehring/tml/download>
  - [14] "View synthesis software manual," *MPEG ISO/IEC JTC1/SC29/WG11*, Sept. 2009, release 3.5.
  - [15] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.
  - [16] S. Jumisko-Pyykkö, "User-centered quality of experience and its evaluation methods for mobile television," *PhD thesis*, Tampere University of Technology, 2011.
  - [17] R. Kennedy, N. Lane, K. Berbaum, and M. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychology*, 3(3), 203–220., 1993.
  - [18] H. Cooligan "Research methods and statistics in psychology," (4th ed.). London: Arrowsmith., 2004.
  - [19] A. Strauss and J. Corbin, "Basics of qualitative research: techniques and procedures for developing grounded theory" (2nd ed.). Thousand Oaks, CA: Sage. 1998.

- [P9] **P. Aflaki**, M. M. Hannuksela, J. Hakala, J. Hakkinen, M. Gabbouj; “Joint Adaptation of Spatial Resolution and Sample Value Quantization for Asymmetric Stereoscopic Video Compression: a Subjective Study, ” International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, September, 2011.

© IEEE, 2011, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# JOINT ADAPTATION OF SPATIAL RESOLUTION AND SAMPLE VALUE QUANTIZATION FOR ASYMMETRIC STEREOSCOPIC VIDEO COMPRESSION: A SUBJECTIVE STUDY

*Payman Aflaki<sup>a</sup>, Miska M. Hannuksela<sup>b</sup>, Jussi Hakala<sup>c</sup>, Jukka Häkkinen<sup>b,c</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

<sup>c</sup>Dept. of Media Technology, Aalto University, School of Science and Technology, Espoo, Finland

## ABSTRACT

A novel asymmetric stereoscopic video coding method is presented in this paper. The proposed coding method is based on uneven sample domain quantization for different views and is typically applied together with a reduction of spatial resolution for one of the views. Any transform-based video compression, such as the Advanced Video Coding (H.264/AVC) standard, can be used with the proposed method. We investigate whether the binocular vision masks the coded views of different types of degradations caused by the proposed method. The paper presents a subjective viewing study, where the proposed compression method is compared with two other coding techniques: full-resolution symmetric and mixed-resolution stereoscopic video coding. We show that the average subjective viewing experience ratings of the proposed method are higher than those of the other tested methods in six out of eight test cases.

**Index Terms**— Low bit-rate video coding, quantization, downsampling, asymmetric stereoscopic video, subjective assessment.

## 1. INTRODUCTION

Asymmetric stereoscopic video is one division of ongoing research for compression improvement in stereoscopic video, where one of the views is sent with high quality, whereas the other view is degraded and hence the bitrate is reduced accordingly. This technique is based on the psycho-visual studies of stereoscopic vision in human visual system (HVS) which demonstrated that the lower quality in a degraded view presented to one eye is masked by the higher quality view presented to the other eye, without affecting the visual perceived quality (binocular suppression theory [1]). The quality difference between the views of a stereoscopic video is commonly achieved by removing spatial, frequency, and temporal redundancies in one view more than in the other. Different types of prediction and quantization of transform-domain prediction residuals are jointly used in many video coding standards. In addition, as coding schemes have a practical limit in the redundancy that can be removed, spatial and temporal sampling frequency as well as the bit depth of samples can be selected in such a manner that the subjective quality is degraded as little as possible.

In [2], a set of subjective tests on a 24" polarized stereoscopic display comparing symmetric full-resolution, quality-asymmetric full-resolution, and mixed-resolution stereoscopic

video coding were presented. The performance of symmetric and quality-asymmetric full-resolution bitstreams was approximately equal. The results showed that in most cases, resolution-asymmetric stereo video with a downsampling ratio of 1/2 along both coordinate axes provided similar quality as symmetric and quality-asymmetric full-resolution stereo video. These results were achieved under the same bitrate constraint.

Objective quality metrics are often able to provide a close approximation of the perceived quality for single-view video. However, in the case of asymmetric stereoscopic video, there are two views with different qualities, and it has been found that objective quality assessment metrics face some ambiguity on how to approximate the perceived quality of asymmetric stereoscopic video [3].

In this paper, we propose a novel compression method for one view of stereoscopic video coding, while the other view is coded conventionally. Our aim is to study the proposed method for asymmetric stereoscopic video due to the fact that it introduces different compression artifacts than those of conventional coding methods and hence the human visual system might mask the coding errors of one view by the other view. Consequently, this paper verifies the assumption that binocular suppression is capable of masking the proposed uneven sample-domain quantization with a systematic subjective comparison of the proposed method with two other compression techniques, namely symmetric and mixed-resolution stereoscopic video coding.

This paper is organized as follows. Section 2 presents the proposed compression method. The test setup and test material are described in Section 3, while Section 4 provides the results. Finally, the paper concludes in Section 5.

## 2. PROPOSED COMPRESSION METHOD

### 2.1 Overview

The proposed encoding approach is depicted in Fig. 1. While the proposed method is applied to the right view in Fig. 1, it can equally be applied to the left view. The proposed coding method consists of the transform-based encoding step for the left view and three steps for the right view: downsampling, quantization of the sample values, and transform-based coding. First, the spatial resolution of the image is reduced by downsampling. The lower spatial resolution makes it possible to use a smaller quantization step in transform coding and hence improves the subjective quality compared to a coding scheme without downsampling. Moreover, downsampling also reduces the computational and memory

resource demands in the subsequent steps. Second, the number of quantization levels for the sample values is reduced using a tone mapping function. Third, transform-based coding, such as H.264/AVC encoding, is applied.

The decoding end consists of the transform-based decoding step for the right view and three respective steps for the left view: transform-based decoding, inverse quantization of sample values, and upsampling. In the first step, the bitstream including coded transform-domain coefficients is decoded to a sample-domain picture. Then, the sample values are rescaled to the original value range. Finally, the image is upsampled to the original resolution i.e. the same resolution as of the left view or to the resolution used for displaying.

In the following sub-section, the key novel parts of the proposed coding scheme, namely the quantization of the sample values in the encoder and their inverse quantization in the decoder are described in details.

## 2.2 Quantization and inverse quantization of sample values

This step of the proposed compression method reduces the number of quantization levels for luma samples. In addition, the original luma sample values are remapped to a compressed range. Hence, the contrast of the input images for transform-based coding and the output images from transform-based decoding is smaller compared to the contrast of the respective original images. The remapping to a compressed value range is typically done towards the zero level, and hence the brightness of the processed images is reduced too.

The proposed method includes the following key steps:

- 1) Before transform-based encoding: reduction of the number of luma quantization levels in the sample domain and scaling of luma sample values to a compressed value range.
- 2) After transform-based decoding: Re-scaling of the decoded sample values in such a way that the original sample value range of the luma sample values is restored.

When the same quantization step size is used for transform coefficients in transform-based encoding, the bitrate of the video where sample values are quantized becomes smaller than that of the same video without sample value quantization. This reduction in bitrate depends on the ratio of the number of luma quantization levels divided by the original number of luma quantization levels, which typically depends on the bit depth. Ratios closer to 0 have very good compression outcome but the quality drop is severe. On the other hand, applying a ratio close to 1 keeps the quality close to the original quality with a smaller relative bitrate reduction. We found ratios greater than or equal to 0.5 to be practical.

The presented sample value quantization operation is lossy,

i.e., it cannot be perfectly inverted, when integer pixel values are in use. Hence, the original pixel values can be only approximately restored by the inverse quantization of sample values.

Based on informal subjective results, the sample value quantization is proposed to be applied only to the luma component. This is because the bitrate saving achieved by quantization of the two chroma components caused a more severe subjective quality reduction than the same bitrate saving achieved by quantizing the luma component more coarsely.

The quantization of sample values can be done in various ways. For example, tone mapping techniques can be exploited [4]. In this paper, linear luma value quantization with rounding is used as expressed as:

$$q = \text{round}\left(\frac{i * w}{2^d}\right) = (i * w + 2^{d-1}) \gg d \quad (1)$$

where:

$q$  is the quantized sample value

$\text{round}$  is a function returning the closest integer

$i$  is the input value of the luma sample

$w$  is the explicit integer weight ranging from 1 to 127

$d$  is the base 2 logarithm of the denominator for weighting

Since Eq. (1) is implemented using integer multiplication, addition, and bit shifting, it is computationally fast. As the sample value range is reduced, the value of  $w$  is required to be smaller than  $2^d$ . With this limitation, Eq. (1) is identical to the formula used for H.264/AVC weighted prediction. The ratio  $(w / 2^d)$  is referred to as the luma value quantization ratio.

Inverse quantization of sample values to their original value range is achieved by:

$$r = \text{round}\left(q' * \frac{2^d}{w}\right) \quad (2)$$

where:

$r$  is the inverse-quantized output value

$q'$  is the scaled value of the luma sample as output by the transform-based decoder

Other parameters are the same values as used in the sample value quantization.

Eq. (2) requires one floating or fixed point multiplication and a conversion of the floating or fixed point result to integer by rounding. If it is preferred to use integer arithmetic in the decoder rather than in the encoder, it is possible to apply Eq. (2) in the encoder and Eq. (1) in the decoder with the condition that  $w$  is greater than  $2^d$ .

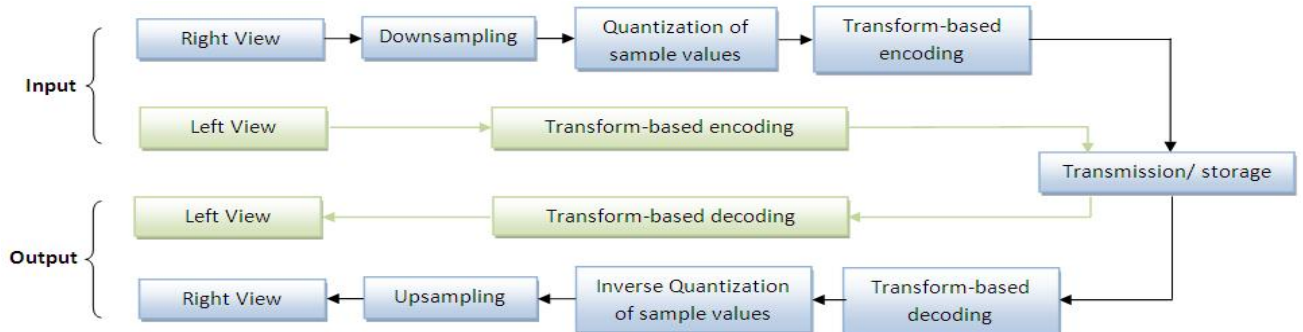


Fig. 1. Diagram of proposed compression method

### 3. TEST SETUP

#### 3.1 Preparation of test stimuli

The subjective assessments were performed with four sequences: Undo dancer, Kendo, Newspaper, and Pantomime. Undo dancer, exemplified in Fig. 2, is a synthetically created photorealistic multiview sequence including a dancing person, reproduced from a motion capture. The other three sequences are common test sequences in the 3D Video (3DV) ad-hoc group of the Moving Picture Expert Group (MPEG). The sequences were downsampled from their original resolutions to the resolutions mentioned in Table 1 in order to be displayed on the used screen without scaling (see Section 3.2). The filters included in the JSVM reference software of the Scalable Video Coding standard were used in this and other subsequent downsampling and upsampling operations [5].

For each sequence, we had the possibility to choose between several camera separations or view selections. This was studied first in a pilot test of 9 subjects. The test procedure of the pilot test was similar to that of the actual test presented in Section 3.2. Several camera views were available for each sequence in the pilot test, and based on the subjective scores achieved, the 4 cm and 5 cm camera separations were chosen for Undo dancer and the rest of test sequences, respectively.

Several bitstreams were coded for each sequence with the following coding methods:

1. Full resolution symmetric stereoscopic video by coding to both views with H.264/AVC. No downsampling or quantization of luma sample values.
2. Mixed resolution stereoscopic video by downsampling the right view and subsequently applying H.264/AVC coding to it while coding the full-resolution left view with H.264/AVC.
3. The proposed coding scheme including downsampling, quantization of luma sample values, and H.264/AVC coding to the right view and coding the left view with H.264/AVC.

The coded left view for each sequence was identical regardless of the coding method. The left view was kept unchanged, because we wanted to assess the perception and acceptability of the left and right eyes presented with different types of quality degradations as caused by transform-domain quantization, spatial downsampling, and sample-domain quantization and to reduce the number of factors which could affect the subjective rating. Joint optimization over both views for the quantization step size for sample values and transform coefficients as well as for the spatial resolution was left to another subjective experiment. As the bitrate of the right view for each



Fig. 2. A frame of Undo dancer sequence

Table 1. Spatial resolution of the right view

	Full	5/6	3/4	1/2
Undo dancer	960x576	800x480	720x432	480x288
Others	768x576	640x480	576x432	384x288

bitstream of the same sequence was kept the same regardless of the coding method used, there was a fair comparison between the coding methods.

In order to have a representative set of options for the second coding method (with downsampling and transform-based coding), three bitstreams per sequence were generated, each processed with a different downsampling ratio for the right view. The subjective results achieved for stereoscopic video in [2] motivated us to use downsampling ratios equal to or greater than 1/2. Hence, downsampling was applied to obtain a spatial resolution of 1/2, 3/4, and 5/6 relative to the original resolution along both coordinate axes. Table 1 presents the spatial resolution of the right view used for different sequences.

As the number of potentially suitable combinations for the downsampling ratio and the luma value quantization ratio is large, their joint impact on the subjective quality was studied first to select particular values for the downsampling ratio and the luma value quantization ratio for the subsequent comparisons between the different coding methods. To reveal potential dependencies at different quantization step sizes for transform coefficients, the bitstreams were generated with several quantization parameter values. Subjective assessment revealed that downsampling ratio 3/4 along with luma value quantization ratio 5/8 tended to provide the best relative subjective results. Thus, these values were consistently used in the subsequent comparisons.

In order to prevent fatigue of test subjects from affecting the test results, only two sets of bitstreams at different bitrates were included in the test. Table 2 presents the selected Quantization Parameter (QP) values for the full-resolution symmetric coding, the resulting bitrates, and the respective average luma PSNR values for the right view of each sequence coded using different coding methods. The PSNR values were derived from the decoded sequences after inverse quantization of sample values and upsampling to the full resolution.

Table 2. Tested bitrates per view, respective QP values per sequence for both higher quality (HQ) and lower quality (LQ), and the respective PSNR values for different coding techniques

		Pantomime	Dancer	Kendo	Newspaper
QP	HQ	41	42	43	42
	LQ	44	45	45	45
Bitrate (Kbps)		445.8	301.5	280.3	148.0
		343.9	224.6	238.5	115.4
Proposed (PSNR-dB)		31.9	29.1	34.1	30.7
		30.6	28.3	33.1	29.5
FR (PSNR)		31.9	29.2	33.3	30.0
		30.0	27.7	32.0	28.3
1/2 (PSNR)		31.7	29.1	35.5	31.7
		30.9	28.3	34.7	30.7
3/4 (PSNR)		32.5	29.5	34.7	31.3
		31.0	28.5	33.5	29.8
5/6 (PSNR)		32.3	29.8	34.1	29.9
		31.0	28.3	32.9	29.2

### 3.2 Test Procedure

12 subjects participated in this experiment of which 7 were women and 5 men. Their age differed from 19 to 32 years with an average of 23.6 years. The candidates were subject to thorough vision screening. Candidates who did not pass the criterion of 20/40 visual acuity with each eye were rejected. All participants had a stereoscopic acuity of 60 arc sec or better. Test clips were displayed on a 24" polarizing stereoscopic screen having the total resolution of 1920×1200 pixels and the resolution of 1920×600 per view when used in the stereoscopic mode. The viewing conditions were kept constant throughout the experiment and in accordance with the sRGB standard [6] ambient white point of D50 and illuminance level of about 200 lux. Viewing distance was set to 93cm which is 3 times the height of the image, as used in some subjective test standards [7].

The subjective test started with a combination of anchoring and training. The extremes of the quality range of the stimuli were shown to familiarize the participants with the test task, the test sequences, and the variation in quality they could expect in the actual tests that followed. The test clips were presented one at a time in a random order and appeared twice in the test session. Each clip was rated independently after its presentation. A scale from 0 to 5 with a step size of 0.5 was used for the rating. The viewers were instructed that 0 means “very bad” or “not natural” and 5 stands for “very good” or “very natural”.

## 4. RESULTS AND DISCUSSIONS

Fig. 3 shows the viewing experience subjective results for all sequences in two different bitrates. Based on the average subjective ratings, it can be seen that the proposed coding method outperformed the other tested coding methods in all cases for the higher bitrate. Furthermore, except for the Dancer sequence, it had similar performance than the best mixed-resolution test case in the lower bitrate. The mixed-resolution coding with 5/6 spatial resolution in the lower quality view outperformed the proposed method for the Dancer sequence at the lower bitrate, while the performance of the proposed method was better than or similar to the performance of the other methods. Moreover, the symmetric full-resolution coding method was clearly inferior to the other tested methods at the lower bitrate.

When comparing the PSNR values presented in Table 2 with the subjective viewing experience results, one can see that PSNR was not representative of the subjective quality in this test.

## 5. CONCLUSIONS AND FUTURE WORK

A novel asymmetric stereoscopic video coding technique was introduced in this paper. The method is based on uneven quantization step size for luma sample values of different views, and it is typically jointly applied with downsampling. The proposed compression method was subjectively compared to full-resolution symmetric stereoscopic video coding and mixed-resolution stereoscopic video coding at different downsampling ratios. The average subjective viewing experience ratings of the proposed method were found to be higher than those of the other tested methods in six out of eight test cases. The results suggest that the human visual system is able to fuse views with different types of quality degradations caused by the proposed method. The provided results should be verified with a greater number of test sequences and more subjective tests to verify these conclusions.

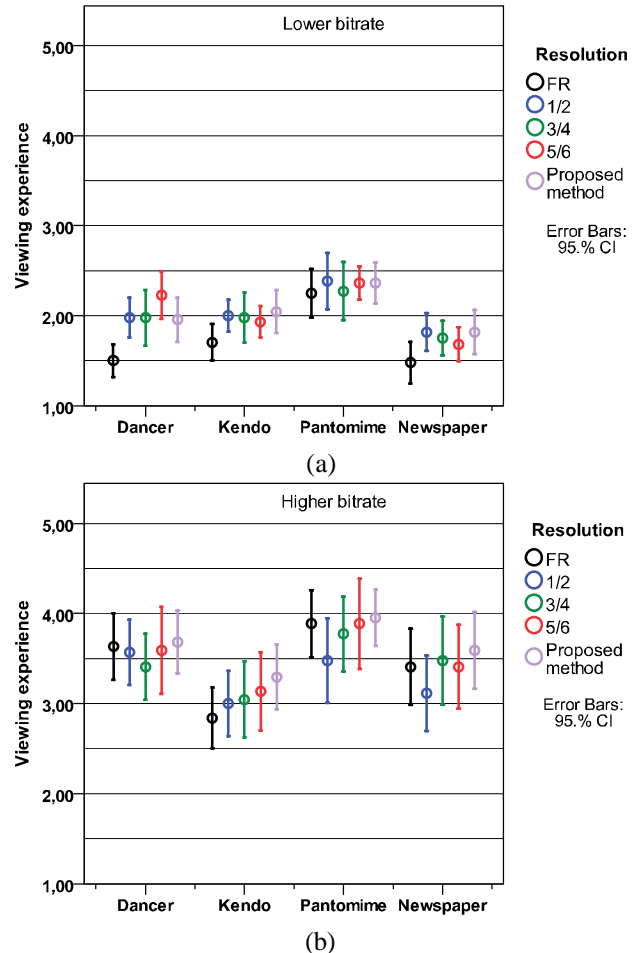


Fig. 3. Results of subjective tests for all sequences using two bitrates (a) lower bitrate (b) higher bitrate

## 6. REFERENCES

- [1] R. Blake, “Threshold conditions for binocular rivalry,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3(2), pp. 251-257, 2001.
- [2] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, “Subjective Study on Compressed Asymmetric Stereoscopic Video,” *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Sep. 2010.
- [3] P.W. Gorley, N.S. Holliman, “Stereoscopic image quality metrics and compression”, *Stereoscopic Displays and Virtual Reality Systems XIX, Proceedings of SPIE-IS&T Electronic Imaging*, SPIE Vol.6803, January 2008
- [4] A. Segall, L. Kerofsky, S. Lei, “New Results with the Tone Mapping SEI Message,” *Joint Video Team, Doc. JVT-U041*, Hangzhou, China, October 2006.
- [5] JSVM Software [http://ip.hhi.de/imagecom\\_G1/savce/downloads/SVC-Reference-Software.htm](http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm)
- [6] B. Girod and N. Färber, “Wireless video,” Chapter 12 in the book “Compressed video over networks,” Marcel Dekker, 2000.
- [7] ITU-T, Subjective assessment methods for image quality in high-definition television, ITU-R BT.710-4



- [P10]** P. Aflaki, M. M. Hannuksela, J. Hakala, J. Hakkinen, M. Gabbouj; “Estimation of subjective quality for mixed-resolution stereoscopic video, ” International 3DTV CONF. Antalya, Turkey, May, 2011.

© IEEE, 2011, Reprinted with permission.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to [http://www.ieee.org/publications\\_standards/publications/rights/rights\\_link.html](http://www.ieee.org/publications_standards/publications/rights/rights_link.html) to learn how to obtain a License from RightsLink.

# ESTIMATION OF SUBJECTIVE QUALITY FOR MIXED-RESOLUTION STEREOSCOPIC VIDEO

*Payman Aflaki<sup>a</sup>, Miska M. Hannuksela<sup>b</sup>, Jussi Hakala<sup>c</sup>, Jukka Häkkinen<sup>b,c</sup>, Moncef Gabbouj<sup>a</sup>*

<sup>a</sup>Department of Signal Processing, Tampere University of Technology, Tampere, Finland;

<sup>b</sup>Nokia Research Center, Tampere, Finland;

<sup>c</sup>Dept. of Media Technology, Aalto University, School of Science and Technology, Espoo, Finland

## ABSTRACT

In mixed-resolution (MR) stereoscopic video, one view is presented with a lower resolution compared with the other one; therefore, a lower bitrate, a reduced computational complexity, and a decrease in memory access bandwidth can be expected in coding. The human visual system is known to fuse left and right views in such a way that the perceptual visual quality is closer to that of the higher-resolution view. In this paper, a subjective assessment of mixed resolution (MR) stereoscopic videos is presented and the results are analyzed and compared with previous subjective tests presented in the literature. Three downsampling ratios 1/2, 3/8, and 1/4 were used to create lower-resolution views. Hence, the lower-resolution view had different spatial resolutions in terms of pixels per degree (PPD) for each downsampling ratio. It was discovered that the subjective viewing experience tended to follow a logarithmic function of the spatial resolution of the lower-resolution view measured in PPD. A similar behavior was also found from the results of an earlier experiment. Thus, the results suggest that the presented logarithmic function characterizes the expected viewing experience of MR stereoscopic video.

**Index Terms**— Video signal processing, video compression, asymmetric stereoscopic video, mixed resolution, subjective evaluation

## 1. INTRODUCTION

Mixed resolution (MR) stereoscopic video compression introduced in [1] is a well-known approach in the field of stereoscopic video coding. In MR stereoscopic video, one view is represented with a lower resolution compared to the other one, while, according to the binocular suppression theory [2], it is assumed that the perceived quality by the Human Visual System (HVS) is closer to that of the higher quality view.

A subjective assessment of full- and mixed- resolution stereoscopic video on a 32-inch polarized stereoscopic display and on a 3.5-inch mobile display was presented in [3]. One of the views was downsampled with ratio 1/2 along both coordinate axes. Uncompressed full-resolution (FR) sequences were preferred in 94% and 63% of the test cases for 32-inch and 3.5-inch displays, respectively. While studying different resolutions for the symmetric stereoscopic video and the higher-resolution view of the MR videos, it was found that the higher the resolution, the smaller the subjective difference was between FR and MR stereoscopic video. The lower resolution view had always a downsampling ratio 1/2 vertically and horizontally.

The study presented in [4] included a subjective evaluation of MR sequences with downsampling ratios 1/2 and 1/4 along both coordinate axes. The results revealed that the subjective image quality of the MR image sequences was preserved well but dropped slightly at downsampling ratio 1/2 and 1/4.

In [5], the impact of downsampling ratio in MR stereoscopic video was studied. Downsampling ratios 1/2, 3/8, and 1/4 were applied vertically and horizontally. A 24-inch polarized display was used with a viewing distance of 70 cm. A correlation comparison between the subjective results and the average luma peak-signal-to-noise (PSNR) showed that there might be a breakdown point between downsampling with ratio 1/2 and 3/8, at which the lower-resolution view became more dominant in the subjective quality. Downsampling ratios 1/2 and 3/8 corresponded to 11.2 and 7.6 pixels per degree (PPD) of viewing angle, respectively. Moreover, it was confirmed that the ocular dominance did not affect the subjective ratings regardless of which view was downsampled in the MR sequences.

In this paper, a subjective test for uncompressed MR stereoscopic video is presented using a test setup similar to but not the same as in [5]. The obtained subjective results are compared to the previous subjective test [5] to see if the above-mentioned breakpoint is valid for a different test setup. Moreover, a novel logarithmic estimation of subjective ratings as a function of PPD values of viewing angle is introduced.

This paper is organized as follows. Section 2 explains the subjective test setup and test procedure. The subjective results are presented and discussed in Section 3. Finally, the paper is concluded in Section 4.

## 2. TEST SETUP

### 2.1 Preparation of the Test Stimuli

Four sequences were used: Pantomime, Dog, Newspaper, and Kendo. They are all common test sequences in the 3D Video (3DV) ad-hoc group of the Moving Picture Expert Group (MPEG). No audio track was used.

For each sequence, we had the possibility to choose between several camera separations or view selections. This was studied first in a pilot test of 9 subjects. The test procedure of the pilot test was similar to that of the actual test presented in Section 2.2. Several camera views were available for each sequence in the pilot test, and based on the subjective scores achieved, the 5 cm camera separation was chosen for all test sequences.

The test clips were prepared as follows. Both left and right view image sequences were first downsampled from their original

**Table 1.** Spatial resolution of sequences

	<i>Full</i>	<i>1/2</i>	<i>3/8</i>	<i>1/4</i>
All sequences	768x576	384x288	288x216	192x144

**Table 2.** Visual angle (in pixels per degree) of the two test setups

Downsampling ratio	Test setup presented in this paper	Test setup presented in [5]
1	30.2	22.8
1/2	15.1	11.4
3/8	11.3	7.6
1/4	7.5	5.7

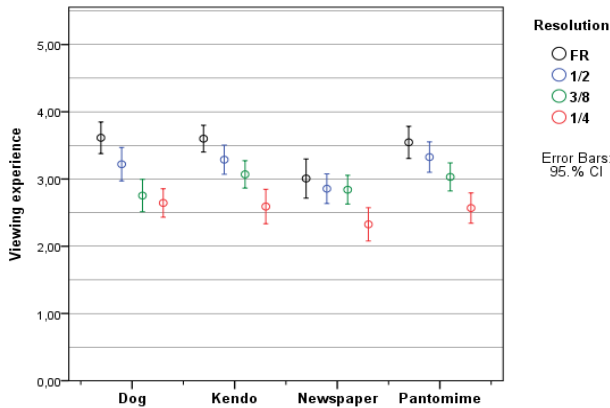
resolution to the “full” resolution mentioned in Table 1. The “full” resolution was selected to occupy as large an area as possible on the used monitor with a reasonable downsampling ratio from the original resolution. As eye dominance was shown to have no impact on which view is provided with a better quality [5], only one set of MR sequences was prepared. The right view was kept in “full” resolution while the left view was downsampled and subsequently upsampled to the “full” resolution. Downsampling ratios 1/2, 3/8, and 1/4 were selected and symmetrically applied along both coordinate axes in order to keep the results easily comparable with those presented in [5]. The filters of the JSVM reference software of the Scalable Video Coding standard were used in the downsampling and upsampling operations [5].

## 2.2 Test Procedure

The same 24” polarizing stereoscopic screen as in [5] was used for subjective experiments. It has width and height of 515 and 322 mm, respectively, a total resolution of 1920x1200 pixels, and a resolution of 1920x600 per view when used in stereoscopic mode.

22 subjects attended this experiment of which 7 were female and 15 were male. The average age of the subjects was 23.5 years. The test viewing distance was changed from 70 cm used in [5] to 93 cm which is 3 times the height of the image, as used in some subjective test standards [7]. Hence, the visual angle differed from that in [5]. Table 2 reports the visual angle in PPD for both test setups.

Prior to the experiment, the candidates were subject to thorough vision screening. Two candidates did not pass the criterion of 20/40 visual acuity with each eye and were thus

**Fig. 1.** Average of viewing experience ratings and the 95% CI

rejected. All participants had a stereoscopic acuity of 60 arc sec or better. The viewing conditions were kept constant throughout the experiment and in accordance with the sRGB standard [8] ambient white point of D50 and illuminance level of about 200 lux.

## 3. RESULTS AND DISCUSSION

### 3.1 Viewing Experience

The average viewing experience ratings and the 95% confidence interval (CI) are presented in Fig. 1. The subjective ratings tend to have less variation in this test than in the test presented in [5]. We observed that 18% and 69% of the total rating interval were covered by the average subjective scores of the sequences in this experiment and in [5], respectively. This result was expected, because increasing the viewing distance diminishes the subjective quality difference among MR stereoscopic videos with different downsampling ratios.

### 3.2 Limit of Downsampling Ratio

With the test setup presented in [5], we found that the downsampling ratio that could be applied before the lower resolution view became dominant in subjective results was between 1/2 and 3/8, i.e., between 7.6 and 11.4 PPD of viewing angle as indicated in Table 2. We studied whether the same PPD ratio threshold appeared in this experiment too. Therefore, as also done in [5], we analyzed the correlation of subjective viewing experience ratings of the presented study with PSNR of the lower resolution view upsampled to the full resolution. Unlike in [5], practically no correlation was found between the subjective viewing experience rating and the average luma PSNR of the lower resolution view for any downsampling ratio. Consequently, the analysis did not reveal the limit of the downsampling ratio for the lower-resolution view in the presented study. We suspect that the lack of correlation could have been caused by the selection of the test sequences and the smaller variation in subjective viewing experience ratings in general. It has also been discovered that the greater the angular size of the display, the more contrast sensitivity the human visual system has [9]. Thus, the threshold angular resolution for mixed-resolution stereoscopic video may also depend on the angular size of the display. As the correlation analysis of the average luma PSNR of the lower resolution view did not lead to conclusions in this test, we explored another approach for discovering the limits of the downsampling ratio in MR stereoscopic video, as presented in the next sub-section.

### 3.3 Logarithmic Estimation of Subjective Ratings

We analyzed ratings achieved in these experiments and those included in [5] against the PPD values of each test setup. A logarithmic relationship was observed between the subjective viewing experience ratings and the corresponding PPD values of each downsampling ratio. The fitting model used to generate the curves in Fig. 2 under the mean square error criterion is as follows:

$$y = C_1 * \log(ppd - k) + C_2 \quad (1)$$

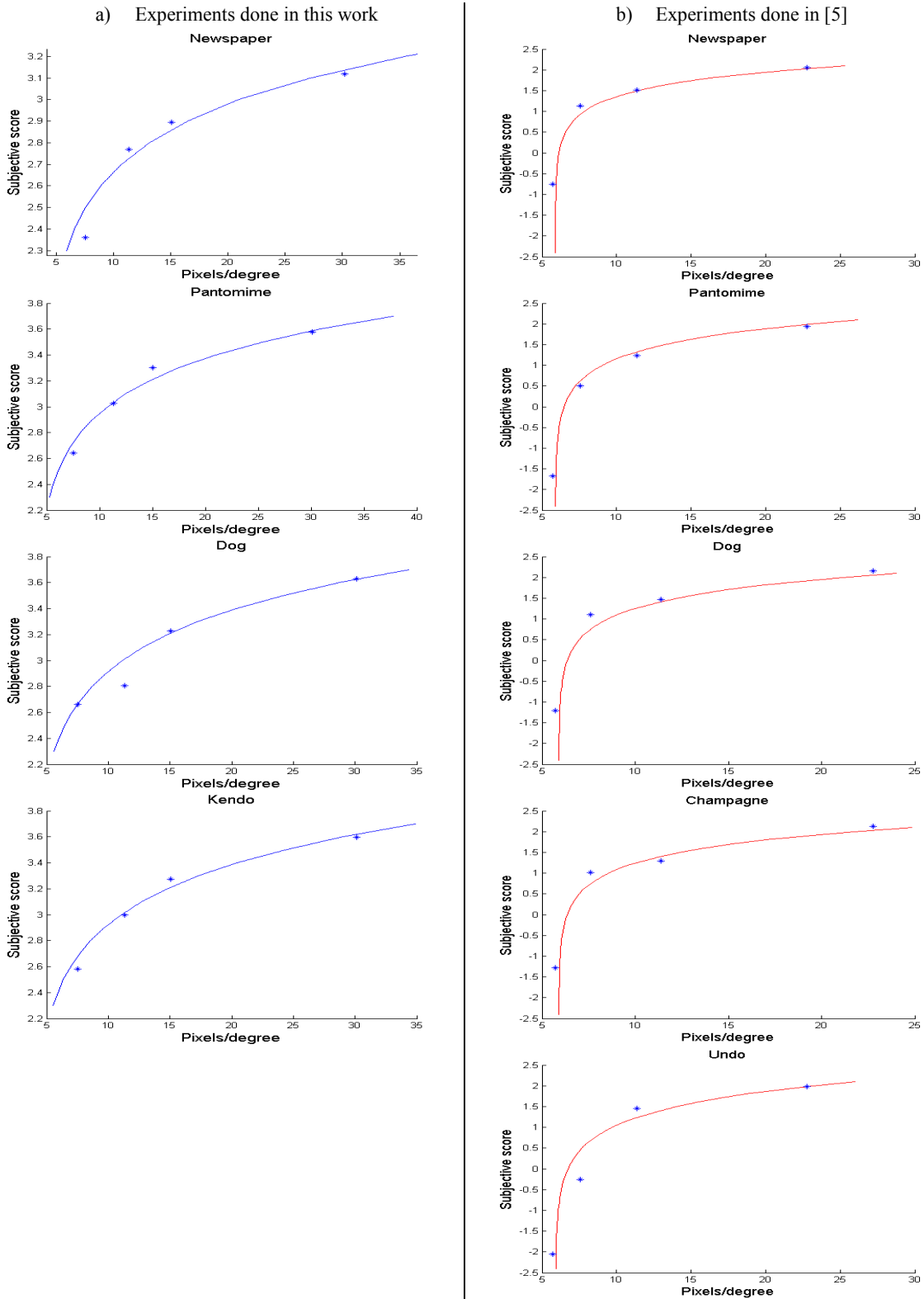
where:

$ppd$  = pixels per degree (PPD) of viewing angle

$C_1, C_2$  = coefficients calculated for each sequence separately

$k$  = fixed offset for each test setup

$y$  = estimated subjective rating



**Fig. 2.** Relation of the subjective average viewing experience ratings and PPD values

Fig. 2 shows the estimated curves for each of the sequences used in this work and in [5]. The subjectively obvious correlation of the data points and the logarithmic estimates were confirmed by

deriving the Pearson correlation coefficients presented in Table 3. Note that as the Pearson correlation measures the linear dependence between two variables, the x-axis of the plots in Fig. 2

should be modified to be  $\log(\text{ppd} - k)$  in order to reflect a correct geometric interpretation of the correlation coefficients in Table 3. On average, the Pearson correlation coefficient between all data points and estimated values among all sequences was 0.97 and 0.98 for tests held in this experiment and [5], respectively.

As the estimation curves turned out to be similar for each test setup, Fig. 3 presents the logarithmic relations estimated in the mean square error sense for all the sequences except Newspaper, whose data points differed significantly from the data points of the other sequences. The other eight test cases fitted the logarithmic estimation very well. The Pearson correlation coefficient between all data points and the joint logarithmic estimation equation is 0.96 for both tests setups in this work and also in [5].

The presented logarithmic equation provided a good estimation of the subjective viewing experience ratings of two different test setups; hence, one could conclude that there might be always a high correlation between MR stereoscopic video subjective scores and the angular resolution of the lower-resolution view measured in PPD. This conclusion should be confirmed by more intensive subjective experiments.

#### 4. CONCLUSIONS

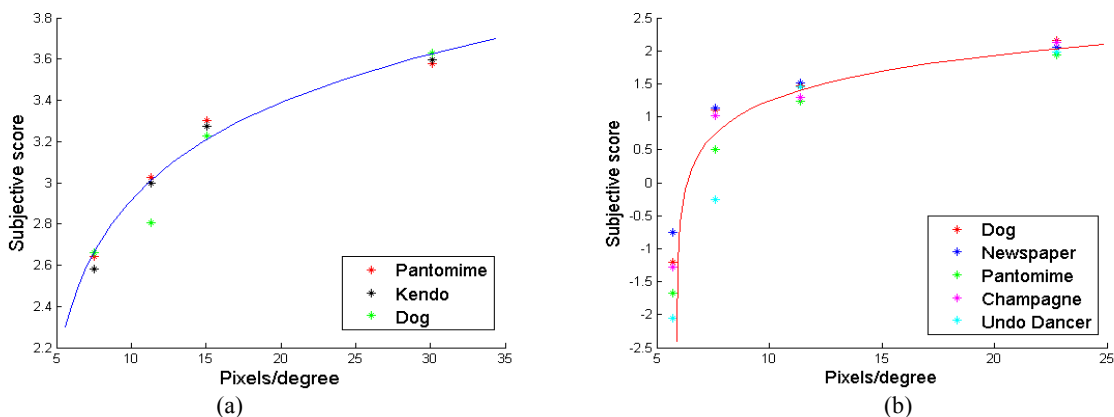
In this work, a set of subjective tests on four asymmetric resolution stereoscopic video sequences was performed. Three different downsampling ratios were applied to the sequences to produce the lower-resolution views. We observed a logarithmic relationship between the subjective viewing experience rating and the angular resolution of the lower-resolution view measured in pixels per degree of viewing angle. The results of the subjective tests presented in this paper and in an earlier work were used to derive two sets of coefficient values for the logarithmic relationship. While the coefficients were remarkably different between the test presented in this paper and the earlier paper, the logarithmic relation provided good estimates of the subjective ratings across all test sequences. Thus, the results suggest that when some subjective evaluations for a few mixed-resolution sequences are available for particular viewing conditions, the proposed logarithmic relation can be used to estimate the subjective rating for other video sequences and downsampling ratios for the lower-resolution view under the same viewing conditions. It is acknowledged that the results should be verified with other video clips and test conditions.

**Table 3.** Pearson correlation coefficient between actual ratings and estimated values, for all sequences of both test cases

Experiment held in this paper		Experiment held in [5]	
Sequence	Pearson Coef.	Sequence	Pearson Coef.
Dog	0.96	Dog	0.97
Newspaper	0.90	Newspaper	0.98
Pantomime	0.97	Pantomime	0.99
Kendo	0.99	Champagne	0.97
		Undo dancer	0.99

#### 5. REFERENCES

- [1] M. G. Perkins, "Data compression of stereopairs," IEEE Transactions on Communications, vol. 40, no. 4, pp. 684-696, Apr. 1992.
- [2] R. Blake, "Threshold conditions for binocular rivalry," Journal of Experimental Psychology: Human Perception and Performance, vol. 3(2), pp. 251-257, 2001.
- [3] H. Brust, A. Smolic, K. Mueller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," Proc. of 3DTV Conference, May 2009.
- [4] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, no. 2, pp. 188-193, Mar. 2000.
- [5] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video", Proc. of 3DTV Conference, June 2010.
- [6] JSVM Software [http://ip.hhi.de/imagecom\\_G1/savce/downloads/SVC-Reference-Software.htm](http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm).
- [7] ITU-T, Subjective assessment methods for image quality in high-definition television, ITU-R BT.710-4.
- [8] M. Anderson, R. Motta, S. Chandrasekar, and M. Stokes, "Proposal for a standard default color space for the Internet – sRGB," Proc. 4th IS and T/SID Color Imaging: Color Science, Systems and Applications, pp. 238-246, Nov. 1996.
- [9] P. G. J. Barten, "The effects of picture size and definition on perceived image quality," IEEE Transactions on Electron Devices, vol. 36, no. 9, pp. 1865-1869, Sep. 1989.



**Fig. 3.** Logarithmic relation for (a) tests done in this work (b) tests done in [5]

Tampereen teknillinen yliopisto  
PL 527  
33101 Tampere

Tampere University of Technology  
P.O.B. 527  
FI-33101 Tampere, Finland

ISBN 978-952-15-3184-2  
ISSN 1459-2045