ROMEO

SEVENTH FRAMEWORK PROGRAMME

Remote Collaborative Real-Time Multimedia Experience over the Future Internet

ROMEO

Grant Agreement Number: 287896

D3.2

Interim Report on QoE and Visual Attention Models

Document description	
Name of document	Interim Report on QoE and visual attention models
Abstract	This document describes the investigations related to Quality of Experience and Visual Attention modelling work that is carried out in the ROMEO project. The purpose of this document is to identify requirement of QoE and VAM for the ROMEO project and to evaluate the performance of existing quality metrics that suit the requirements. This work will have important implications on content processing and network optimization tasks of ROMEO.
Document identifier	D3.2
Document class	Deliverable
Version	3.0
Author(s)	V. De Silva, C. Kim, I. Elfitri, N. Haddad, S.Dogan (US); I. Politis, T. Kordelas, M. Blatsas (UPAT); P. tho Pesch, H. Qureshi (IRT); G. Dosso (VITEC);
QAT team	Didier Doyen (TEC), Jonathan Rodriguez (IT), Nikolai Daskalov (MMS)
Date of creation	03-Sep-2012
Date of last modification	27-Sep-2012
Status	Final
Destination	European Commission
WP number	WP3
Dissemination Level	Public
Deliverable Nature	Report

TABLE OF CONTENTS

TABLE OF CONTENTS 3

LIST OF FIGURES 5

LIST OF TABLES 7

1 INTRODUCTION 8

1.1 Purpose of the Document 8
1.2 Scope of the Work 8
1.3 Objectives and Achievements 8
1.4 Structure of the Document 8

2 QOE AND VAM MEASUREMENT FRAMEWORK OF ROMEO 9

2.1 Introduction 9
2.2 QoE/VAM Requirements of ROMEO project 9
2.3 State-of-the-art on QoE of Compressed Sterescopic Video 10
2.3.1 Assessment of Asymmetrically coded stereoscopic video 10
2.3.2 Quality Metrics for measuring compression artefacts 11
- PSNR-HVS optimized for Stereoscopic video (PHVS-3D) 11
- Local Disparity Distortion weighted SSIM (SSIM-Ddl) 11
- Compressed Stereoscopic Video Quality Metric (CSVQ) 12

2.4 State-of-the-art of Visual Attention Modelling 12

3 QOE MODELLING OF COMPRESSION ARTEFACTS IN STEREOSCOPIC VIDEO 14

3.1 Introduction 14
3.2 Psycho-physical Experiments on Compression Artefacts 14
3.2.1 Analysis Of Tolerance Levels of Interocular Blur Suppression (IBS) 14
- Experimental Setup 14
- Effect of spatial frequency on IBS tolerance 16
- Effect of luminance contrast on IBS tolerance 16

3.2.2 Comparison of inter-ocular blur suppression and compression artefact suppression 16
- Experimental Setup 16
- Subjective Results 16

3.2.3 Discussion of Subjective Results 17
- Comparison of asymmetric blurring and asymmetric coding of stereoscopic images 17

3.3 Subjective Assessment of Compression Artefacts 18
3.3.1 Experimental Setup 18
3.3.2 Subjective Results 18
3.3.3 Discussion of Subjective Results 19
3.4 Metrics for Measurement of Compression Artefacts in stereoscopic video 19
3.5 Proposed initial metric to measure stereoscopic video quality 21

4 QOE MODELLING OF PACKET LOSS ARTEFACTS IN STEREOSCOPIC VIDEO 23

4.1 Networking Aspects of QoE 23
4.1.1 The Impact of Packet Loss 23
4.1.2 The Impact of Physical Layer Errors 24
4.2 The Proposed QoE Model 25
4.3 Experimental Setup 25
4.4 Objective and Subjective Results 27
4.5 Analysis of Subjective results 28

5 MEASUREMENT OF RENDERING ARTEFACTS FOR MULTIVIEW VIDEO APPLICATIONS 30

5.1 Introduction 30
5.2 Metrics/Software for measuring the quality of depth maps for rendering 30
5.3 Objective Results and Discussion 32

5.4 Subjective Assessment of Rendering Artefacts 34
5.5 Future work on depth/disparity map quality evaluation 34
6 MEASUREMENT QOE OF SPATIAL AUDIO 35
6.1 Introduction 35
6.1.1 Spatial audio QoE attributes 35
6.1.2 Acoustical parameters for QoE prediction 36
6.2 Subjective Test for Spatial Audio QoE Prediction 37
6.2.1 Experiment design 37
6.2.2 Test environment and procedure 38
6.2.3 Listening test results and findings 39
6.3 Comparison with SoA in audio QoE and new avenues for prediction model refinement 42
6.4 Comparison of audio codec performance for spatial audio compression 42
6.5 Future work related to audio QoE 43
7 VISUAL ATTENTION MODELLING FOR 3D VIDEO 45
7.1 Introduction 45
7.2 Saliency Detection 45
7.2.1 Visual salience based on the spatial information 45
7.2.2 Temporal Image Signature 46
7.2.3 Depth map based saliency 47
7.2.4 Feature weighting 48
7.3 Content Format Adaptation 49
7.3.1 Smooth Motion 50
7.3.2 3D constraint 51
7.3.3 Cropping parameter definition 51
7.4 Evaluation and demonstrator 51
8 CONCLUSIONS AND FUTURE WORK 52
9 REFERENCES 53
APPENDIX A: GLOSSARY OF ABBREVIATIONS 56

LIST OF FIGURES

Figure 1: Functional processes of ROMEO and required QoE measurements required at each process 10

Figure 2: Block diagram for PHVS-3D 11

Figure 3: Block diagram for SSIM-Ddl 12

Figure 4: Block diagram for CSVQ 12

Figure 5: Examples of Test Stimuli 14

Figure 6: Subjective results for IBS tolerances for 16 subjects. 15

Figure 7: Comparison of objective measurement of blur at just noticeable point in quantization and Gaussian blurring 18

Figure 8: DMOS vs. Different QP Combinations 20

Figure 9: Correlation plots for MOS vs. Metrics 21

Figure 10: Correlation against different weights 21

Figure 11: Performance of the proposed metric 22

Figure 12: Aspects affecting QoE 23

Figure 13: UEP scheme with two priority blocks 24

Figure 14: Test-bed platform 26

Figure 15: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model (“Martial Art” sequence) 27

Figure 16: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model (“Munich” sequence) 28

Figure 17: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model (“Panel Discussion” sequence) 28

Figure 18: System architecture of Depth Map Quality measurement 30

Figure 19: Holes caused by disoccluded regions. (a) Cause regions (b) Virtual left view of ‘Ballet’ sequence (white pixel are the holes) 31

Figure 20: DMQ Results for rendering with single disparity/depth map 33

Figure 21: DMQ Results for double disparity/depth map 34

Figure 22: User interface for subjective tests on Audio QoE. On each page there are 6 processed stimuli, 1 hidden reference and 2 anchors 38

Figure 23: Means and 95% confidence intervals of QoE scores for different angular deviation of auditory scene from the video. 40

Figure 24: Scatter plot of the subjective QoE scores averaged across the subjects, against the predicted QoE values using the results of ridge regression. 41

Figure 25: Performance of several audio codecs at various bitrates[47] 43

Figure 26: Performance of AAC multichannel and MPEG Surround in terms of ODG score 43

Figure 27: Most attractive features for human brain. From left to right: color, orientation and intensity. .. 45

Figure 28: Overview of suggested approach..... 46

Figure 29: Overview of Temporal Image Signature. Upper diagram: Saliency detection applied on a single image. Lower diagram: Saliency detection applied on multiple frames..... 47

Figure 30: Steps from captured depth map (b) to MB based ROI partitioning (d) 48

Figure 31: First diagram version of proposed system. 49

Figure 32: Suggested cropping concept..... 50

LIST OF TABLES

Table 1: Subjective Comparison of Just noticeable level of Gaussian Blurring and Quantization 17

Table 2: Bit Rate reductions achieved by the two asymmetric processing schemes 17

Table 3: QP Combinations of test sequences 19

Table 4: Performance of Existing Metrics 20

Table 5: Encoding and Streaming Parameters 26

Table 6: Result of correlation analysis using SPSS between the QoE score and the magnitude of angular deviation of auditory scene from the presented video. 39

Table 7: Result of ridge regression using SPSS, after iterations to reduce the number of independent variables. 41

1 INTRODUCTION

1.1 Purpose of the Document

The work on Quality of Experience (QoE) and Visual Attention modelling (VAM) is an integral part of work package 3 (WP3), which would find applications in 3D media compression and rendering. The purpose of D3.2: “Interim Report on QoE and visual attention models” is to report the work carried out so far in the area of QoE and VAM.

1.2 Scope of the Work

This deliverable reports the initial work carried out towards the development to QoE metrics to measure artefacts in stereoscopic media that occur due to rendering, compression and transmission related issues such as packet losses. In the context of VAM, this deliverable reports the investigations and a new algorithm for salient region detection in stereoscopic video. Furthermore, this deliverable also presents a review on the state-of-the-art for audio QoE measurement.

1.3 Objectives and Achievements

The main objectives of QoE and VAM research in ROMEO are listed as follows:

- Provide appropriate QoE assessment techniques for compressed 3D media.
- The QoE work will be supported by research into new 3D visual attention modelling techniques.
- Definition of novel objective evaluation metrics designed to assess the perceived quality of spatial audio and 3D free-viewpoint video.

Towards reaching the above main objectives, the following are the main achievements related to QoE and VAM during the first year of ROMEO.

- In line with Milestone 9 (MS9) the subjective experiments are completed to evaluate an initial QoE model.
- Initial work on assessing and developing Visual attention models suitable for stereoscopic video have been performed and reported in this deliverable.

1.4 Structure of the Document

This deliverable is structured as follows: Section 2 presents the QoE and VAM framework of ROMEO project and its requirements. Section 3 describes the work on QoE modelling for compression and section 4 describes the work on QoE modelling of packet loss artefacts on user perception. Section 5 and 6 presents quality metric development process for rendering artefact measurements and QoE aspects of Audio, respectively. VAM work is reported in Section 7 and section 8 concludes this deliverable with some insights to future work.

2 QOE AND VAM MEASUREMENT FRAMEWORK OF ROMEO

2.1 Introduction

The ROMEO project focuses on the delivery of live and collaborative 3D immersive media on a next generation converged network infrastructure. The concept of ROMEO will facilitate application scenarios such as immersive social TV and high quality immersive and real-time collaboration. In order to support these application scenarios, the ROMEO project envisage to develop new methods for the compression and delivery of 3D multiview video and spatial audio, as well as optimising the networking and compression jointly. The solution proposed by ROMEO is to combine the DVB-T2 and DVB-NGH broadcast access network technologies together with a QoE aware Peer-to-Peer (P2P) distribution system that operates over wired and wireless links.

The P2P technology has the potential to provide a more cost effective and flexible delivery solution for future 3D entertainment services. However, the P2P distribution can create problems for network operators, by consuming significant amounts of bandwidth. Many ISPs have begun to use bandwidth throttling during peak demand periods to constrain P2P applications. This can create serious problems for real-time delivery of multimedia content. ROMEO aims to produce a content aware P2P overlay, which will be able to scale the quality of the multimedia data in response to bandwidth constraints and congestion problems. To achieve this in an optimal fashion, new 3D Visual Attention Models (VAM) and 3D Quality of Experience (QoE) models will be developed and used within the scope of ROMEO.

In this section we present the envisaged QoE framework that would be used in ROMEO. In this discussion, the QoE development work within ROMEO are categorized in to three main areas; i.e. QoE modelling for 3D video, for spatial audio and modelling of networking aspects. The developments in QoE modelling for 3D video include subjective assessment and modelling of compression artefacts, rendering artefacts in view point synthesis and visual comfort factors. The main contributions of QoE work on audio aspects relate to the effects of listening point mismatches, audio-video synchronization losses. Finally, the Networking aspects related to QoE such as the effect of variation of delay and packet losses due to congestion and fading will also be discussed in this paper.

Rest of this section is organized as follows. In section 2.2, we describe the overall ROMEO architecture and the requirements of QoE measurement. The section 2.3 describes the relevant work in the area of QoE, section 2.4 describes the related work in the area of VAM.

2.2 QoE/VAM Requirements of ROMEO project

This section briefly describes the main building blocks of the overall ROMEO architecture and the requirements for a QoE Measurement Framework.

The ROMEO project aims to deliver 3D multiview video synchronously on both DVB and P2P networks. To guarantee Quality of Service (QoS), all the users will be transmitted a stereoscopic video pair over the DVB network. For users with a good network conditions, additional views of the same scene will be transmitted through P2P streaming. For optimal performance of the system, the DVB and P2P stream would need to be synchronized. The scenes captured by multi camera rigs and spatial audio microphones will be compressed using a multiple description scalable video codec and an analysis-by-synthesis based spatial audio codec. Finally, audio and video stream need to be synchronized to play back the content.

The Figure 1 illustrates the functional block diagram of the ROMEO project architecture. At each of the blocks we have identified specific factors of QoE that is related to the block. During content capturing it is important to consider factors such as visual comfort of the captured content and the sensation of depth. For example, depth/disparity variations in the scene should not exceed certain thresholds, and scene changes should be planned such that they do not cause visual discomfort. Compression of video with the aid of state-of-the-art codecs yield artefacts such as blurring and blocking. The effect of such artefacts should be carefully modelled to preserve subjective quality. With the approach of Visual Attention Modelling video

encoding can take into account what part of the visual scene probably draws the viewers' attention. This information helps to improve subjective quality. Issues such as error concealment due to packet losses occurring due to congestion or fading also need to be considered during QoE modelling. To cater user requirements, such as arbitrary view point switching and multiview rendering, the intermediate views need to be synthesized (i.e. rendered) from the available views. Depending on the quality of disparity map and the whole filling algorithm utilized, the rendered views will have different artefacts that affect the user perception. During audio rendering it is also important to measure listening point and viewpoint mismatch and its effect on the overall 3D perception.

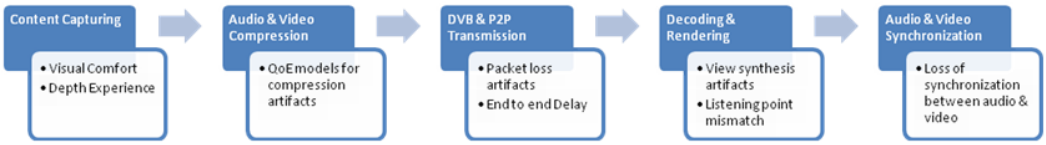


Figure 1: Functional processes of ROMEO and required QoE measurements required at each process

The above mentioned QoE factors need to be monitored and measured and used for decision making within the scope of the ROMEO project. According to the Description of Work (DoW) of ROMEO, the main development work in the area of QoE will be on modelling QoE for compression/packet loss/view synthesis artefacts and issues related to audio visual synchronization.

2.3 State-of-the-art on QoE of Compressed Sterescopic Video

This section describes related work found in literature that are of significance to the QoE contributions of ROMEO.

2.3.1 Assessment of Asymmetrically coded stereoscopic video

One of the major challenges faced in the attempt to deploy advanced 3D Video applications is the high bandwidth required to transmit multiple views simultaneously. One solution to this problem in the case of stereoscopic 3D is asymmetric coding, which makes use of a phenomenon known as "Binocular Suppression". Accordingly, when one stereoscopic view is encoded at a higher quality and the other view is encoded at a slightly lower quality, the perceived subjective visual quality is dominated by the higher quality view.

There are several forms of Binocular Suppression. If the two eyes are provided with similar images, but of unequal contrast the perception of the Human Visual System (HVS) is dominated by the higher contrast image. This process is known as Interocular Blur Suppression (IBS). Julesz [1] explained the binocular suppression phenomenon with the aid of the experiments he performed with random dot stereograms. According to Julesz, when the low or high frequency (or both) components of the binocular stimuli are identical binocular fusion will arise. On the contrary if the frequency components are different, another form of binocular suppression known as binocular rivalry will occur. In the case of binocular rivalry either one image of the stereo pair is seen or both images are seen alternately.

In Ref. [2], Perkins et al. theoretically analyzed the stereoscopic image compression problem by way of rate-distortion theory. He proposed mixed resolution coding, where the resolution of one image is reduced while the other is kept at its original high resolution. To the best of our knowledge, the first and the only attempt to use the psycho-physical findings of binocular suppression to develop an asymmetric stereoscopic image coder is found in [3]. In Ref. [3] authors use the findings of Liu and Schor [4] regarding the binocular suppression zone to develop a wavelet based encoder that eliminates redundant frequency information from the stereoscopic image pair.

Recently, there have been significant efforts towards identifying the limits for the level of asymmetry or the quality difference with which the stereoscopic images can be compressed [5] [6] [7]. In Ref. [5] the bounds of asymmetric stereoscopic video compression and its relationship to eye-dominance are examined by way of a user study.

In Ref. [8] by way of subjective experiments authors suggest that when one of the stereoscopic view pair is encoded at a sufficiently high quality (i.e. a PSNR of about 40dB), the other view can be encoded at a lower quality above a display dependent threshold without subjective visual quality degradation . This lower quality threshold or the just noticeable level of asymmetry is around 31dB for a parallax barrier display and 33dB for a polarized projection display.

2.3.2 Quality Metrics for measuring compression artefacts

In the recent past there have been several attempts to develop metrics to measure the effect of compression artefacts in stereoscopic video. This section outlines several of those efforts and describes three of those metrics in detail.

- **PSNR-HVS optimized for Stereoscopic video (PHVS-3D)**

The PSNR metric is optimized considering luminance masking functions of the Human Visual System (HVS) in [9], known as PSNR-HVS. In [10] PSNR-HVS was further improved by incorporating DCT coefficient masking, which became known as PSNR-HVSM. The authors in [11] developed a new metric for assessment of compressed stereoscopic video in mobile devices by considering the Mean Squared Error (MSE) of the 3-dimensional DCT (3D-DCT) calculated on the similar/corresponding blocks in the stereoscopic views. In doing so the metric considers the binocular vision by combining together the left and right corresponding blocks and also model the pseudo-random movements the eyes are performing while processing spatial information know as “saccades”. A block diagram of the PHVS-3D metric is illustrated in Figure 2.

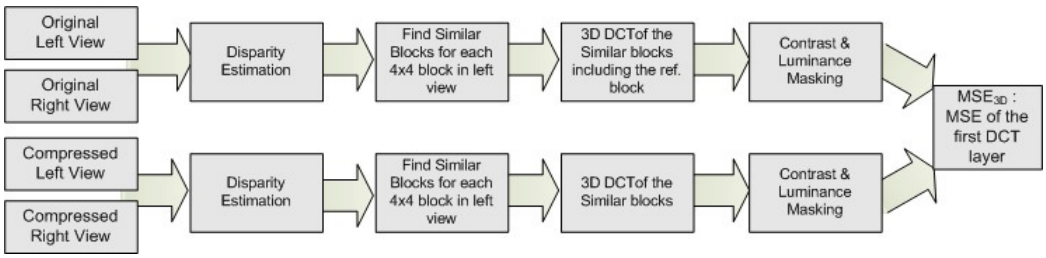


Figure 2: Block diagram for PHVS-3D

- **Local Disparity Distortion weighted SSIM (SSIM-DdI)**

In this metric the SSIM metric is modified to account for the disparity distortion due to compression/processing [12]. This metric is an image quality assessment metric. The SSIM map of left and right images are weighted by the disparity distortion map as illustrated in Figure 3. The disparity distortion map is obtained by calculating the per-pixel Euclidean distance between the original disparity map (generated by estimating disparity with original stereoscopic pair) and distorted disparity map (generated from the compressed stereoscopic pair).

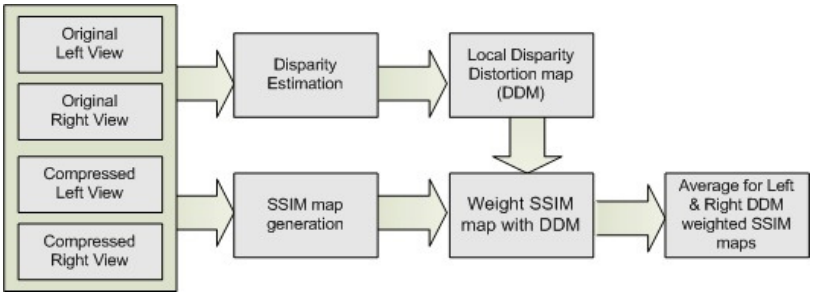


Figure 3: Block diagram for SSIM-Ddl

- **Compressed Stereoscopic Video Quality Metric (CSVQ)**

This is the most recently published metric for measuring compression artefacts in stereoscopic videos [13]. As illustrated in Figure 4 CSVQ metric considers three features, namely the blur, blocking artefacts and similarity of the compressed stereoscopic views. Similarity between the compressed stereoscopic views is measured by considering the similarity of confidently corresponding pixels. For this purpose the original disparity map (generated from the original stereoscopic pair) is utilized.

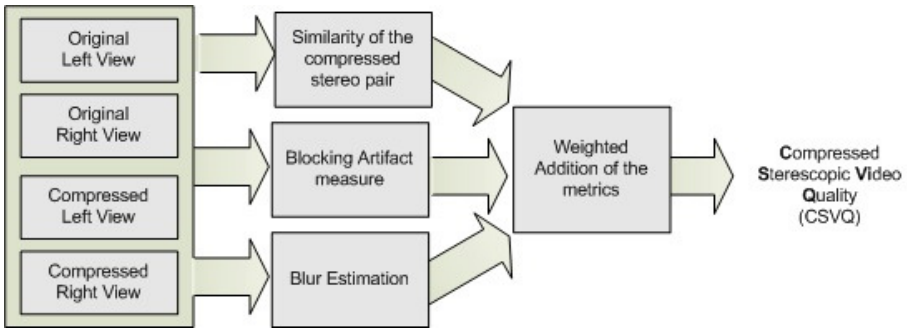


Figure 4: Block diagram for CSVQ

2.4 State-of-the-art of Visual Attention Modelling

Saliency detection can be divided into two types. Image based saliency detection is applied on single images where the visual scene is analyzed. Those methods focus on finding salient regions that stand out from the background. Video based saliency detection is applied on videos and aims to find salient motion in a video. Both type of saliency detection methods on images and videos has been studied in recent years.

The early idea of visual Attention Modeling comes from human visual system, where the fast but simple pre-attentive process of detection is the first stage of human vision. Models for guessing the position of distinct features within the scene are strongly inspired by two models regarding visual perception. One of the most influential models is Feature Integration Theory (FIT) [14]. An extension to FIT model is Guided Search (GS) [15]. Both try to explain schematically a human's behavior when he is looking at a visual scene. The idea of automatic saliency detection is also comparable to these models and often used as inspiration.

Many approaches for saliency detection originate from the field of communications engineering. Generally methods based on fast Fourier transformation offer a fast way to process information and it is often capable of real time processing. A relatively robust method for visual saliency detection is introduced by Huo and Zhang [16]. Their Spectral Residual Approach (SR) relies on the observation that log spectra of different images share similar information. Huo and Zhang assume that the average image has a smooth spectrum and any object or area that causes an aberration from the smooth spectrum will catch the viewers' attention. In order to detect salient regions the image is first transformed into frequency domain. Then the spectrum is smoothed and subtracted from the original one. The result is transformed back into the time domain.

Later, Cui et al. [17] extend the SR approach to temporal domain. A technique introduced by Cui et al. 'Temporal Spectral Residual' (TSR) [17] deals with the detection of salient motion by extracting salient features from the video sequence. Principally TSR approach applies SR approach on temporal slices (XT and YT planes). The XT and YT are the planes of image lines in a temporal domain. The resulting image only comprises the unusual parts of the image's spectrum, which are considered to represent saliency with motion information.

Other investigators, Ma et al. [14] and Achanta et al. [18], [19] estimate saliency using center surround feature distance and maximum symmetric surround method [20]. Ming-Ming Cheng et al. proposed regional contrast based saliency extraction [21]. Furthermore, predicting where humans look has proven to be important information for many application areas such as object-of-interest segmentation [22][23], motion detection, frame rate up-conversion [24], [25] and image re-targeting [20]. Lately, the idea of detecting salient region by applying a quaternion DCT and quaternion signatures to calculate the visual saliency [49] was presented and used for the application of face detection.

Another approach is presented by Vu and Chandler [26]. They focus more on the question how advantages of different methods can be combined best. Thus they use a rating system to process different saliency maps. First a set of well-known feature extraction methods like sharpness, color and luminance distance or contrast is defined. All methods are applied to the input image and propose a feature salient map each. Vu and Chandler argue that the key to robust object detection is a rating mechanism that weights all feature salient maps. They suggest using cluster density as a weighting decision. The final salient map is a combination of the weighted maps.

A similar approach is presented by Christof Koch and Shimon Ullman [27]. The paper proposes that visual features that contribute to attentive selection of a stimulus could be combined on a single map: the Saliency Map. This map integrates the normalized information from the individual feature map into one global measure of conspicuity. This theory is already today a base and a reference in this field of research.

Recently Hou et al. proposed a method called Image Signature (IS) [28], which define the saliency using the inverse Discrete Cosine Transform (DCT) of the signs in the cosine spectrum. IS approach discards amplitude information across the whole frequency spectrum without introducing visual distortion, thus keeping only the sign of each DCT component.

3 QOE MODELLING OF COMPRESSION ARTEFACTS IN STEREOSCOPIC VIDEO

3.1 Introduction

Modeling the effects of compression artifacts on perceived 3D video quality is a major research area within the ROMEO QoE tasks as well as in the general research communities. This section of the deliverable outlines the research work carried out to measure the effects of compression artifacts. Firstly, this section describes the psychophysical experiments performed to subjectively assess various thresholds of stereoscopic perception and secondly, we describe the subjective experiments performed to assess the effects of compression artifacts in stereoscopic video.

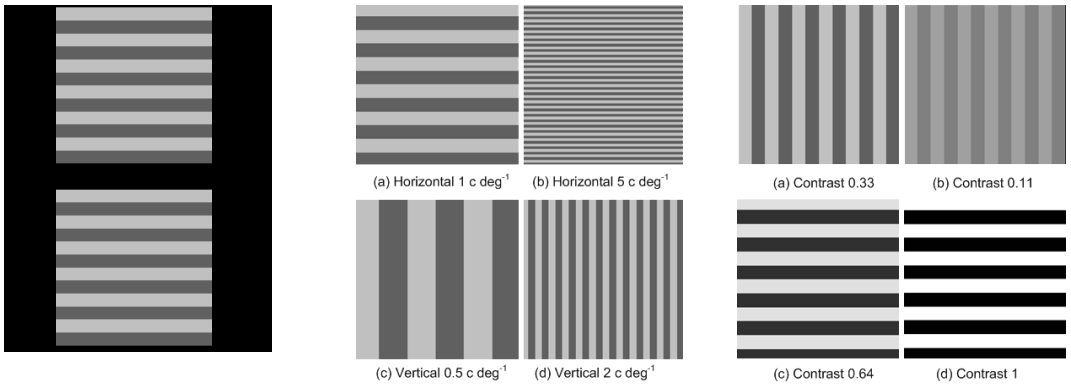
3.2 Psycho-physical Experiments on Compression Artefacts

In an asymmetric coding scenario, Most of the asymmetric coding techniques discussed in section 2.3.1 are based on subjective experiments performed under different conditions and sequences with different characteristics. Unfortunately, most of the techniques discussed earlier do not explicitly consider the psycho-physical phenomenon underlying asymmetric coding. Besides, the phenomenon of interocular blur suppression has been investigated especially in mono-vision correction related studies [29] that cannot be directly used in asymmetric coding. The compression artefacts introduce both blurring and blocking artefacts in to the stereo views, which are perceived differently by the human visual system. To address the above issues, a set of psycho-physical experiments are performed to identify the thresholds of interocular blur suppression.

We measure the just noticeable level of asymmetric blur at various spatial frequencies, luminance contrasts and orientations. This work was published in [30]. This section describes the psychophysical experiments in detail.

3.2.1 Analysis Of Tolerance Levels of Interocular Blur Suppression (IBS)

This section describes the psycho-physical experiments that are carried out to measure the tolerance levels of IBS. The first experiment investigates the variation of tolerance with varying spatial frequency. The objective of the second experiment is to analyze the variation of IBS tolerance with varying contrast levels. The results obtained from the experiments are also discussed within this section.



(a) An example of a test stimulus as viewed by the subjects

(b) Examples of test stimuli used to investigate the effect of spatial frequency

(c) Examples of test stimuli used to investigate the effect of luminance contrast

Figure 5: Examples of Test Stimuli

- **Experimental Setup**

The main test stimulus is a pair of square wave gratings as shown in Figure 5(a). The stimulus as given in Figure 5(a) is constituted by a stereoscopic image pair. The top square wave

gratings are kept unchanged, while one of view of the stereo pair that constitutes the bottom square wave gratings is gradually blurred using a Gaussian low pass filter. The standard deviation of the Gaussian filter (σ) is increased in steps of 0.1 per every second. The spatial frequency and the contrast of the stimuli are kept unchanged throughout each reading. The experiments are performed at different spatial frequencies and contrast levels. The subjects will indicate when they perceive a difference in the bottom gratings with relative to the top gratings.

When displayed on the screen, each of the gratings (as shown in Figure 5(a)) is a square area with each side measuring 24 cm. When observed at a distance of 2.3m from the screen, a set of gratings corresponds to a visual angle of 6^{o}. The spatial frequency of a stimulus is measured by the number of cycles per visual angle (c deg^{-1}). In other words, how many times the luminance values alternate within a visual angle of one degree.

The Michelson contrast γ as given in Eq. (1) is used to define luminance contrast of stimuli. In Eq.(1), L_{max} and L_{min} refer to the maximum and minimum luminance levels present in the stimuli.

$$\gamma = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \qquad (1)$$

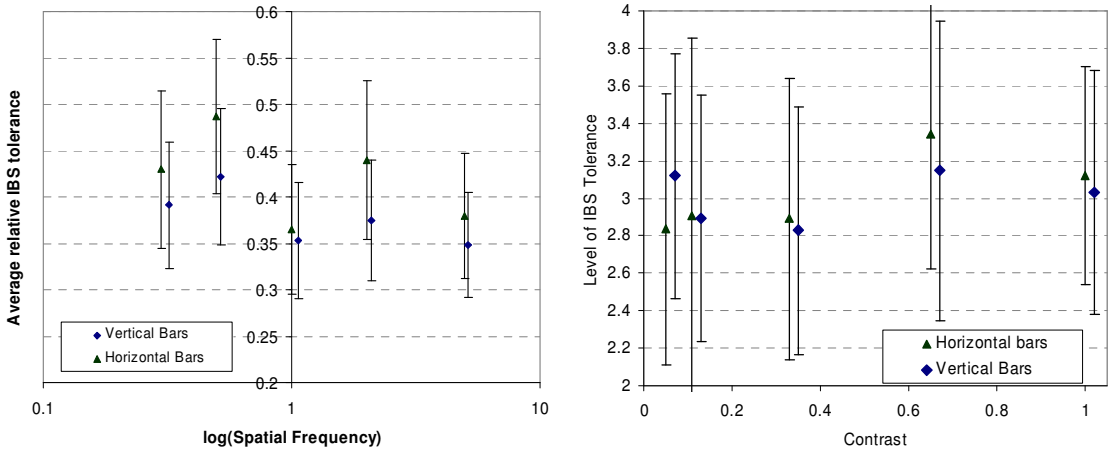
All the lighting in the test room is turned off, and the ambient illumination is measured at 5lux.

The tolerance level of IBS is presented as the maximum level of blur that could be tolerated. The standard deviation of the Gaussian filter (σ) at the maximum tolerable level of blur is used as the measure of IBS tolerance.

To illustrate the relative variation of the tolerance with different frequencies/contrasts the level of IBS tolerance of an individual j is normalized as given in Eq. (2).

$$r_{i,j} = \frac{o_{i,j} - o_{\max,j}}{o_{\max,j} - o_{\min,j}} \qquad (2)$$

In Eq. (2), o_{ij} refers to the IBS tolerance of individual j to a stimuli i and $o_{max,j}$ and $o_{min,j}$ refers to the maximum and minimum values of subject j for all the stimuli in the particular experiment. Thus, $r_{i,j}$ refers to the relative tolerance of an individual i to a particular stimuli j .



(a) IBS vs. Spatial Frequency

(b) IBS vs. Luminance Contrast

Figure 6: Subjective results for IBS tolerances for 16 subjects.

- **Effect of spatial frequency on IBS tolerance**

This experiment investigates whether the tolerance levels of IBS is affected by the spatial frequency of the content.

The tolerance level is measured at spatial frequencies of 0.3, 0.5, 1, 2 and 5 c deg^{-1}. The tolerance is measured for both horizontal and vertical gratings at each frequency. At a time the bottom gratings of one view of the stereoscopic image pair is gradually blurred and the experiment is repeated by blurring the other view in a similar way. Thus, there are a total of 20 stimuli used in this experiment. The contrast is kept constant for each of the stimuli at 0.3. The Figure 5(b) illustrates few of the stimuli used in this experiment.

The Figure 6(a) summarizes the average relative tolerance of IBS at different spatial frequencies for vertical and horizontal gratings. In general, subjects can tolerate more blur in horizontal spatial frequencies than vertical frequencies. The psycho-physical tolerance level, in terms of σ, across different frequencies varies between 3.3 and 2.8, which is a relatively low variation, considering the width (3 σ) of the filter.

- **Effect of luminance contrast on IBS tolerance**

This experiment investigates whether the tolerance levels of IBS is affected by the luminance contrast of the content.

The tolerance level is measured at luminance contrasts of 0.05, 0.11, 0.33, 0.64 and 1. As in previous experiment there are a total of 20 stimuli used in this experiment corresponding to the five contrast levels utilized. The spatial frequency is kept constant for each of the stimuli at 1 c deg^{-1}. The Figure 5(c) illustrates few of the stimuli used in this experiment.

The subjective results are summarized in Figure 6(b). Similar to the variation of IBS tolerance at different frequencies, the vertical frequencies have low tolerance than horizontal frequencies. However, at very low contrasts (<0.1) tolerance of IBS in vertical frequencies is much higher than horizontal frequencies.

3.2.2 Comparison of inter-ocular blur suppression and compression artefact suppression

This section describes the subjective experiment that is carried out to compare and contrast the effect of blurring and quantization processes towards asymmetric stereoscopic image perception.

- **Experimental Setup**

In this experiment subjects evaluate 8 asymmetric stereoscopic image pairs, 4 of whose asymmetry is achieved by blurring and the other 4 of whose asymmetry is achieved by compression. 16 male subjects, aged 23-40, with normal or corrected visual acuity, participate in this experiment. 4 images of resolution 1920x1080 are displayed on the same JVC display discussed in section 2.1. The illumination of the viewing environment is 20lux (very dark).

For compression we use the Intra frame encoder of the H.264 Joint Model (JM) reference software version 15.1, the quantization parameter is increased at steps of 1. For blurring we use a Gaussian low pass filter of 30 pixels wide, and the standard deviation is increased at steps of 0.1. At the beginning of each sequence, subjects see a symmetric stereoscopic image, i.e. the left and right images are both the uncompressed image. Then the subjects navigate forward of the video sequence, one frame at a time until they notice a difference in quality from the beginning. At each step, either the quantization parameter (QP) or the standard deviation (Stdev) of the Gaussian filter applied on the right view is increased, while the left view is kept unchanged. At the just noticed level of perceived difference, the software records the frame number, where this frame number corresponds either to the QP or to Stdev of the right view at the just noticeable perception difference.

- **Subjective Results**

The results of this subjective experiment are summarized in terms of QP or Stdev of the Gaussian filter at the point of just noticeable perception difference, and the corresponding

PSNR and the bit rate of the right view. To provide an indicative bit rate, in the case of Gaussian blurring, the right view at the point of just noticeable difference (JND) is encoded using the JM Intra coder to a PSNR of approximately 40dB. The PSNR of 40dB is selected as the benchmark quality achievable by the JM Intra coder that does not yield visually discernible artefacts, which is the same value used in [7] to encode the high quality view of the asymmetric pair. The results of this experiment are summarized in Table 1.

Table 1: Subjective Comparison of Just noticeable level of Gaussian Blurring and Quantization

Image	Quantization			Gaussian Blurring		
	QP	PSNR (dB)	Bit rate (kbps)	Stdv	PSNR (dB)	Bit rate (kbps)
Badminton	43	29.44	5159	5.9	16.67	3963
Beergarden	39	30.96	7017	5.4	22.52	2288
Cafe	40	36.65	1692	5.6	26.27	1022
GT Fly	32	38.74	4420	5.3	28.36	561

The results presented in Table 1 indicate that higher level of asymmetry in terms of PSNR difference or bit rate difference between the stereoscopic pair could be achieved by Gaussian blurring.

3.2.3 Discussion of Subjective Results

- **Comparison of asymmetric blurring and asymmetric coding of stereoscopic images**

Image or video compression techniques yield in several types of artefacts, such as blurring, blocking and ringing artefacts. Due to the quantization of high frequency components, there is blurring of the image. Furthermore, due to the block based architecture of standardized video codecs, such as H.264/AVC, there are blocking artefacts, which appear as edges, especially in areas of low spatial frequencies.

The results of the subjective experiment described earlier further highlights the difference between blurring and quantization. To illustrate the difference of the two asymmetric processing schemes we summarize the bit rate reductions that are achievable in *Table 2*. Accordingly, it is clear that achieving stereoscopic asymmetry in terms of blurring is more effective in terms of subjective quality.

Table 2: Bit Rate reductions achieved by the two asymmetric processing schemes

Image	Left View Bit rate (kbps) for 40dB	Right View			
		Quantization		Gaussian Blurring	
		Bit rate (kbps)	ΔBR%	Bit rate (kbps)	ΔBR%
Badminton	20321	5159	74.61	3963	80.5
Beergarden	28768	7017	75.61	2288	92
Cafe	3472	1692	51.26	1022	70.6
GT Fly	5471	4420	19.21	561	89.7

The level of blur at the point of just noticed difference, objectively measured in terms of Average Edge Width (AEW) [31], in the two cases is illustrated in Figure 7. It is clear that amount of blur present, when asymmetry is identified in asymmetric compression is very much lower than the case of asymmetric blurring. The high standard deviations for case of Gaussian blurring indicate that just noticeable point with asymmetric blurring varied significantly among individuals. However, in the case of quantization, most of the viewers agreed on the point of just noticeable difference, which was mainly identified by the visibility of blocking artefacts.

Most of the recently reported subjective assessments [5] [6] [7] were carried out by encoding the video at various quantization parameters and varying the level of quantization until subjects perceive a difference. However, as described above, the video coding artefacts

consists of two parts, blurring and blocking artefacts. The psycho-physical experiment reported in this paper suggests that humans can tolerate a significant amount of asymmetric blur before perceiving a difference. This level of tolerance is much higher than the level of blur that is present in just noticeable asymmetrically encoded video sequences [7]. Thus, the low tolerance of asymmetric encoding as compared to asymmetric blurring is mainly attributed to the blocking artefacts.

The human visual system (HVS) can successfully perceive high spatial frequencies of a stereoscopic image pair, masking any blur. Blocking artifacts introduce high spatial frequencies in to the stereoscopic pairs, whereas Gaussian blurring reduces high spatial frequencies. Therefore, the effect of blocking artefacts is perceived in stereoscopic viewing, but the effect of blurring is not perceived. If there are blocking artefacts in a certain area of one view of a stereo pair, and if they are not present in the corresponding area of the other view, this gives rise to binocular rivalry as described in the introduction. It is reasonable to assume that subjects identify the level of asymmetric encoding by way of blocking artefacts, rather than by the asymmetric level of blur.

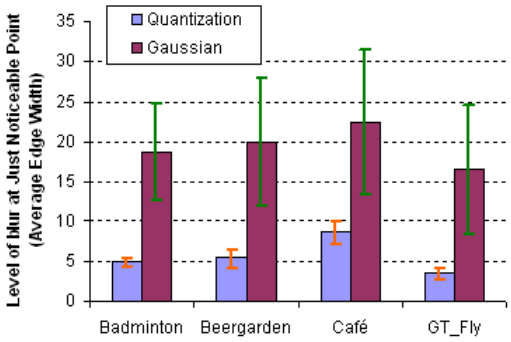


Figure 7: Comparison of objective measurement of blur at just noticeable point in quantization and Gaussian blurring

3.3 Subjective Assessment of Compression Artefacts

This section describes the subjective experiments performed to analyze compression artefacts in stereoscopic video.

3.3.1 Experimental Setup

For these assessments we encode 6 High Definition stereoscopic video sequences, including 3 videos captured within the scope of ROMEO project. Since sterescopic content will be transmitted based on the frame compatible mode, each view is down sampled to 960x1080 before encoding. Three of the videos are encoded using the H.264/SVC reference software (JSVM v9.19) and three others are encoded using High Efficiency Video Coding (HEVC) reference software (HM v8.0). The JSVM and HEVC encoded sequences were tested in two separate subjective experiments. 200 frames of each video sequence are encoded, corresponding to an 8 second clip. The selection quantization parameters (QPs) for left and right view of test sequences are summarized in Table 3.

16 non-expert male subjects attend the subjective experiment, where the sequences are viewed on 46 inch JVC, passive stereoscopic display. The light level of experiment room is 200lux. The subjects have normal or corrected-to-normal visual acuity measured with the Snellen eye-chart. The subjects are aged between 21 and 40 with a mean age of 31 years.

3.3.2 Subjective Results

This section reports the subjective results obtained for the test sequences listed above. The subjects were asked to rate the perceived stereoscopic video quality compared to the reference, on a scale of 1-100. The subjective scores are then normalized to 0-1 and the difference between the original and the processed video is calculated for each test stimuli, for

each subject. The difference score is averaged for each test stimuli to calculate the Differential Mean Opinion Scores (DMOS) value. The Figure 8 illustrates the obtained DMOS for various QP combinations considered for left and right views of the stereoscopic videos. These subjective scores will be used to develop the QoE metric to measure quality of compressed stereoscopic video. The various combinations used for the subjective experiments are selected find a balance between the number of test stimuli used in the experiment and the richness of the subjective score database that will be eventually used for training of the QoE metric.

Table 3: QP Combinations of test sequences

Sequence	*Encoder*	*QP combinations*
Band	JSVM	(12,12), (20,20), (32,32), (40,40), (12,20), (32,12), (12,40)
Martial Arts	JSVM	(16,16), (24,24), (16,24), (32,16), (16,44), (32,24), (44,32)
Panel Discussion	JSVM	(36,36), (48,48), (24,16), (16,36), (48,16), (24,36), (48,24)
Orchestra	HEVC	(15,15),(25,25), (35,15), (35,25), (45,15), (45,25), (45,35)
Beergarden	HEVC	(35,35),(45,45), (35,15), (35,25), (45,15), (45,25), (45,35)
Race	HEVC	(15,15),(25,25), (15,25), (15,35), (15,45), (25,35), (25,45)

3.3.3 Discussion of Subjective Results

Several important observations can be made out of the subjective results illustrated in Figure 8. These are outlined as follows.

- There is a clear difference in DMOS scores between sequence combinations that include a view encoded at QP≥40, and those which does not include one.
- Except for two test stimuli of Orchestra (35,15) and Band (32,32), the DMOS value is less than or equal to 0.22.
- The main difference observed between H.264 encoded videos and HEVC encoded videos is that, DMOS for the HEVC encoded videos are either “Excellent” or “poor”, where as the DMOS for H264 encoded videos show more spread of opinion between “Excellent” and “poor”.

A comprehensive statistical analysis of the subjective results will be presented in the next deliverable on QoE, where we will check for significant differences between DMOS values for different test stimuli.

3.4 Metrics for Measurement of Compression Artefacts in stereoscopic video

In this section we analyze the performance of existing metrics in measuring the quality of compressed stereoscopic video. For this purpose we use two perceptual 2D video quality metrics (VQM & SSIM) and an existing stereoscopic video quality metric ($SSIM_{Ddl}$). In the case of 2D quality metrics the average of left and right views is used for analysis. In particular we analyse the correlation of the metrics with the subjective scores.

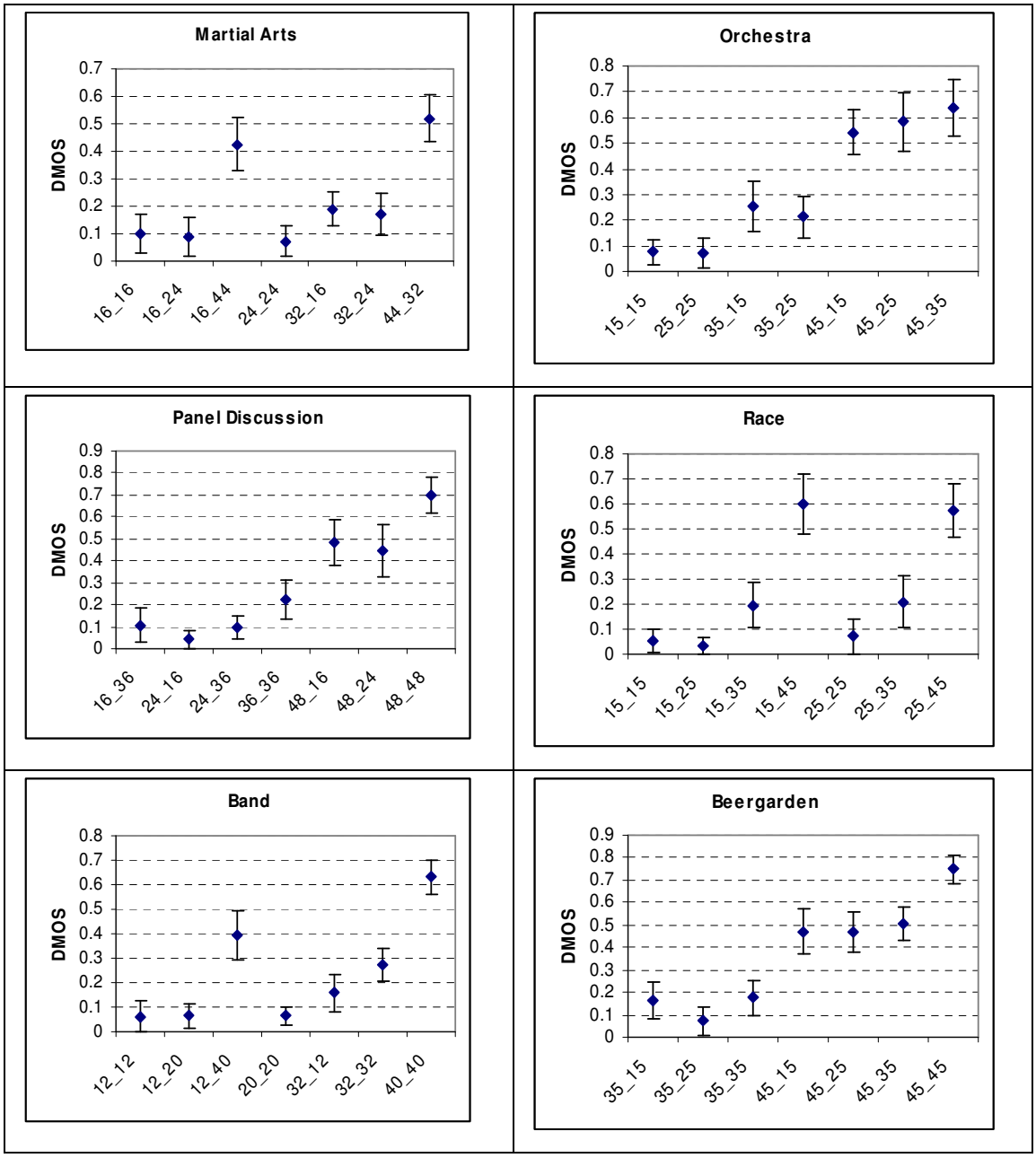


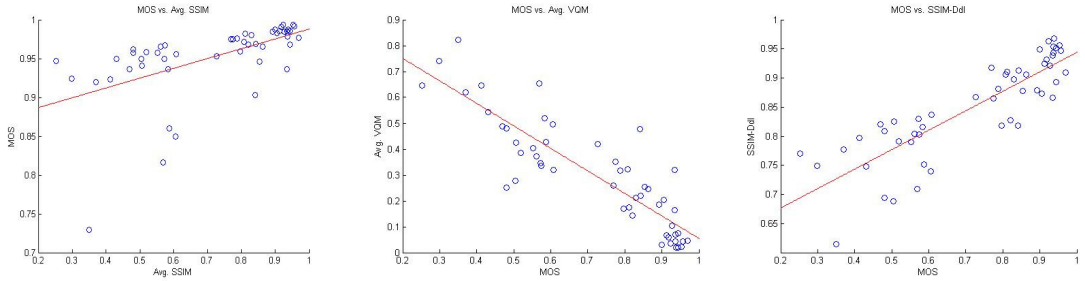
Figure 8: DMOS vs. Different QP Combinations

Table 4: Performance of Existing Metrics

Metric	*Pearson's Correlation*	*RMSE*
Average SSIM	0.552	0.0406
Average VQM	**0.871**	0.104
SSIM-Ddl	0.848	0.0444

The Table 4 summarizes the performance of the existing metrics in terms of their capability to predict the perceptual quality of stereoscopic video. The Table 4 provides the correlation

coefficient of the linear regression (Pearson's Correlation) and the corresponding root mean squared error (RMSE). The correlation plots of the investigated metrics against the MOS are illustrated in Figure 9. According to these results, Average VQM is the most suitable of the considered metrics.



(a) MOS vs. SSIM *(b) MOS vs. VQM* *(c) MOS vs. SSIM-DdI*

Figure 9: Correlation plots for MOS vs. Metrics

3.5 Proposed initial metric to measure stereoscopic video quality

None of the investigated models perform satisfactorily to predict the MOS. In this section we present results for the proposed initial QoE metric. In this method we take into account the binocular suppression theory of the HVS. Accordingly we take in to account two features: maximum VQM of the stereoscopic pair and minimum VQM of the stereoscopic pair. We weight the maximum VQM and minimum VQM over a range of values to find the best linear fit. The weights at the combination of weights that give the highest R^2 are taken as the best weights. For the training purpose we use exactly half of the subjective data set.

By taking the maximum VQM, we assume that binocular perception is mostly driven by the high quality image. However, as described in ref. [7] when the low quality view is below a certain threshold the perception is driven by the low quality view. Therefore the metric should also consider the minimum VQM as a feature for binocular combination. The correlation at different weights for these two features is illustrated in Figure 10. Accordingly the proposed metric is given as;

$$VQM_{Stereo} = 0.375 \cdot \max(VQM_{left}, VQM_{right}) + 0.125 \cdot \min(VQM_{left}, VQM_{right}) \quad \text{Eq. 3.5.1}$$

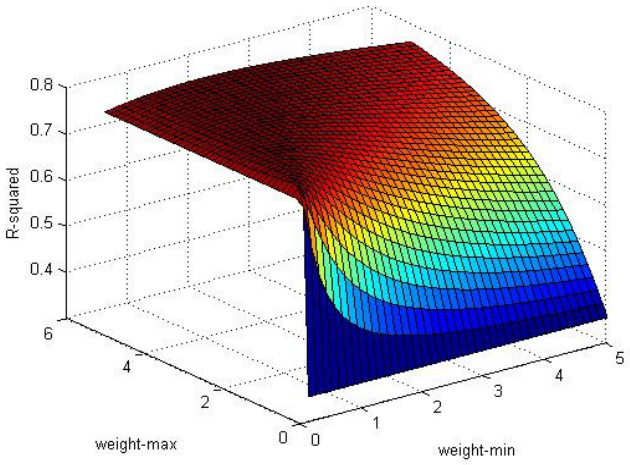


Figure 10: Correlation against different weights

A similar approach was also considered for SSIM. The performance was better than using the average SSIM, but it was quite inferior compared to the VQM performance. Correlation plot of the proposed metric with DMOS is illustrated in Figure 11. The overall performance parameters of the proposed metric are calculated by considering the entire subjective data set. The proposed initial metric performs better than the average VQM in terms of the correlation as well as the RMSE. The performance parameters are as follows:

- R-Squared: 0.787
- Pearson's Correlation: 0.887
- RMSE: 0.0551

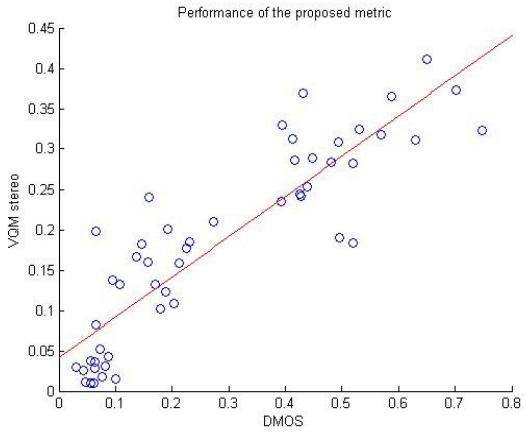


Figure 11: Performance of the proposed metric

The proposed metric is a simplistic extension of the VQM metric. While the performance of it is better than the investigated metrics, there is still a long way to achieve a metric that predicts the MOS well. Another important aspect of the proposed metric is that it does not include disparity estimation, which is less time consuming. Furthermore, the disparity estimation techniques are yet to reach a satisfactory performance level suitable for incorporation in quality assessment metrics. In the next year of ROMEO we will investigate the performance of other stereo metrics as well as perform more subjective experiments to validate/train the QoE models.

4 QOE MODELLING OF PACKET LOSS ARTEFACTS IN STEREOSCOPIC VIDEO

As it has been already stated 3D video quality comprises a variety of perceptual attributes, including overall image quality, naturalness, presence, perceived depth, comfort, immersiveness, etc. This section provides an analysis of the work in progress for developing and investigating an objective QoE model for 3D video streaming that takes into account network parameters (packet losses) and physical layer errors (noise, interference, fading, etc.) In the preliminary study that is included in this section the proposed model is based on the relationship between objective and subjective video quality measures. The investigation is concerned with scalable encoded (using JSVM SVC encoder) stereoscopic video (left-and-right views), although future work will also include depth maps, in accordance to the ROMEO encoding scheme.

4.1 Networking Aspects of QoE

There is a real challenge in devising models which accurately perform learning and statistical functions to model service behaviours by taking into account parameters such as, arrival pattern request, service time distributions, I/O system behaviours, and network usage. These attempts aim to estimate QoE for both resource-centric (utilization, availability, reliability, incentive) and user-centric (response time and fairness) environments over certain predefined QoS/QoE thresholds.

Correlation between QoS and QoE has yet to be defined, although there were several attempts to relate the human perception of quality to the objective network conditions. The aforementioned relationship seems definitely non-linear [32]. In [33] network delivery speed and latency proved to be important to human satisfaction, however only bandwidth and latency time in an integrated network environment are not sufficient aspects to represent the whole spectrum of issues that render a service non-appealing to the end-user.

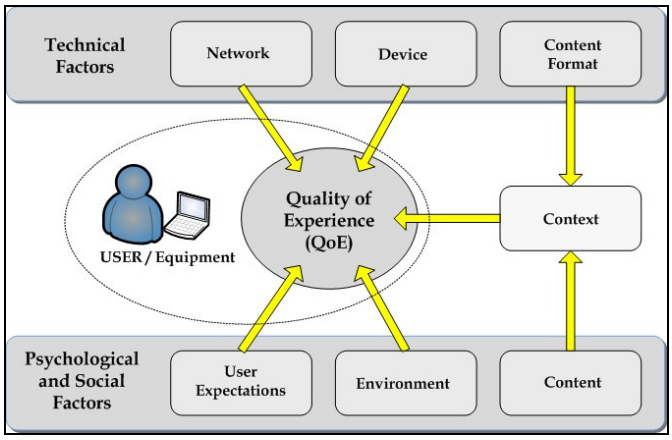


Figure 12: Aspects affecting QoE

4.1.1 The Impact of Packet Loss

Packet loss is considered one of the most influential network parameters regarding QoE, since it directly affects the quality of service presented to the end user regardless if this is video, image, voice or plain text. Information alteration due to packet loss is strongly related to network congestion, but also encompasses the effects of all degradation introduced by media coding. Therefore perceived quality is directly connected to several aspects of the overall transmission topology, with packet loss being the dominant one. This networking technology can clearly be categorized as provider-oriented, enabling network operators to guarantee content delivery in a predictable way. Nevertheless, QoE is mostly based on the actual feelings of users, something that packet loss calculating techniques fail to estimate. Real-user feedback has been the only method which clearly demonstrates this connection, although such surveys are difficult to be conducted in an economical way and most important, in real-

time [34]. In Figure 12 the factors that have a strong impact on QoE are presented. It becomes clear that Network parameters are definitely important, but not unique into a QoE estimation model. Psychological and Social factors along with Contextual elements play a vital role therefore cannot be neglected in any way, nevertheless these factors need to be further investigated and their impact on the overall QoE is part of future work within ROMEO.

4.1.2 The Impact of Physical Layer Errors

The wireless channel is characterized by error bursts of varying size and random occurrences. These errors are the result of noise, interferences, fading, etc., which are present in the wireless medium and characterize its quality. The aim is to transmit efficiently high rate error-delay sensitive video data over error-prone environments. Using Unequal Error Protection (UEP) for video transmission in an erroneous environment leads to better performance for the overall video quality. UEP provides different kind of protection to different parts of video data with different level of importance. Furthermore, it protects the most significant data from any error incidents, in order to improve the video quality, without increasing the size of video data [35]. Thus, the less significant video data can be more sensitive to bit errors, due to the different level of protection. In ROMEO a UEP scheme that is directly related to the channel conditions, will be used at the last hop of video transmission.

In order to use an UEP method, first of all video data must be divided into different layers based on their importance. In the case of two layers, the most significant information corresponds to High Priority (HP) data, as the less significant information to Low Priority (LP) data. Usually, the HP data can be decoded by themselves with acceptable video quality. On the other hand, the LP data are used to improve the video quality. So errors occurred in LP data doesn't have destructive consequences to the video decoding, while errors occurred in HP data lead to large degradation of video quality. Hence, UEP achieves better video quality by providing strong protection to the most significant information.

UEP provides different kind of protection to parts of video data that have different level of importance. As part of this preliminary study, we propose utilize a UEP method, which considers *L*-levels of protection and an equal number of blocks of source symbols *B1,B2,…, BL* which correspond to each level. Thus, one block B_i includes symbols that have higher importance than those included in B_{i+1}. According to the UEP scheme these blocks are duplicated according to a number of Repeat Factors (RF_i), where *i=1,…,L*. Therefore, a virtual source block is created, as shown in an example in Figure 13. Two UEP strategies are considered:

1. High Priority to I-frames, where I-frames, of both views, are considered important information and are grouped in blocks of high priority.
2. High Priority to Left view, where the left view is considered important information and is given priority against the right view.

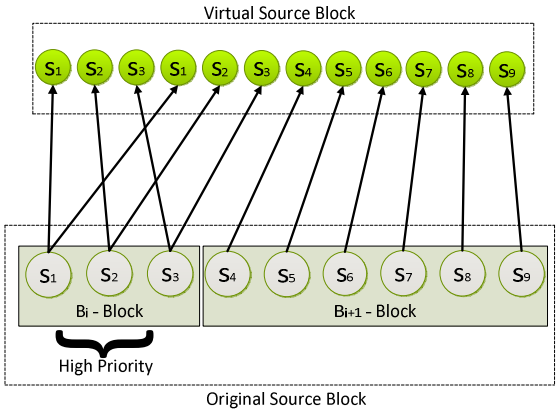


Figure 13: UEP scheme with two priority blocks

4.2 The Proposed QoE Model

The proposed QoE model is based on an objective quality metric for stereoscopic video, which can be used to assess the perceived quality affected by video coding and network impairments, during video transmission. The user's QoE is determined by both the network QoS, which depends on transport parameters such as latency, jitter and packet loss rates and the video compression distortion. In order to achieve high user QoE during real time applications, the video transmission and compression components need to be constantly monitored and controlled. Since this study is preliminary and in order to minimize the computational complexity, the proposed metric is based on the Video Quality Metric (VQM) of both left and right views and correlates closely with subjectively measured MOS. Hence, it can effectively replace subjective evaluation tests during real time video applications.

The subjective quality assessment of a 3D video transmitted over IP network produces MOS ratings that are influenced by the distortion of both left and right views due to compression distortion and packet losses. In the context of this study, both views are encoded and transmitted independently of each other over identical networks (i.e. identical networking conditions, packet loss, latency, etc.).

Therefore, it is safe to assume that the proposed objective QoE metric (Obj_QoE) is a linear combination of the objectively measured VQM values of the left and right views. However, this metric needs also to reflect other aspects which are the essence of 3D QoE, including visual fatigue and perception of depth, naturalness, presence, etc. Such parameters will improve the overall experience of the viewer of the rendered sequence. Finally, the implementation of error protection and correction mechanisms during the video transmission will also improve the overall perceptual quality of the stereoscopic video. Based on these assumptions the proposed metric can be expressed as follows:

$$Obj_{QoE} = A\left(\frac{w \cdot VQM_l + z \cdot VQM_r}{2}\right)$$

$$and\ A \in (0,1) \qquad (w,z) > 0$$

Where, VQM_l and VQM_r are the subjectively obtained VQM values of the left and right view respectively. The parameter *A* depends on the content of the video, the user's Visual Human System, as well as, other physiological and physical factors that contribute on how a user perceives video artefacts on the received video, due to transmission losses. This parameter is assumed to improve the perceived QoE of the 3D video, hence has a positive value. Finally, the weight factors *w*, *z* are UEP dependent and based on the UEP strategy that is selected each time, may minimize the distortion effect on one of the two views. In the context of this study these weight factors are related with each other according to the following:

- *w=z,* in case of using High Priority to I-frames UEP strategy.
- *w<z,* in case of using High Priority to Left-view UEP strategy.

In the above QoE estimation model and for the purposes of this study the value of the weight factors is: *A=0.322*, *w=z=0.702*. These values have been estimated during experimentations and correspond to the particular sequences. There is a stereoscopic effect along with the human perception, which although it cannot be quantified at this point, it has been inserted in the formula as parameter A. A is a weight factor that takes positive values less than 1, thus it actually increases QoE_Obj metric, because we assume at this point that the viewer has better visual perception of the 3D video compared to 2D. It is clear that further study is required in order to render the above model capable of estimating the QoE of any video sequence.

4.3 Experimental Setup

In order to validate the proposed model three left-right 3D video sequences with different motion characteristics, have been selected for the experiments. The first sequence is "Martial

Art", which is a high motion sequence with texture variation and standing camera, while the second and the third sequences named "Munich" and "Panel Discussion" are low motion sequence with low texture variation. All three sequences have high resolution 1920x1080 pixels and frame rate 25 frames/sec. Both video sequences are encoded using the SVC JSVM reference software based on the IPPP... sequence format, an intra-frame period of 8 frames and Context Adaptive Binary Arithmetic Coding (CABAC). Both views of the sequences are encoded in three quality layers using Medium Grain Scale (MGS) mode using the same set of quantization parameters (QP=36, 30, 24).

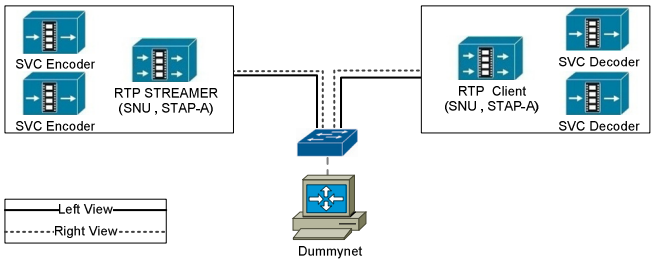


Figure 14: Test-bed platform

Furthermore, the test-bed includes an RTP streamer/packetizer, which is developed capable of creating RTP packets. The packetizer is able of creating RTP packets using the Single NAL Unit payload structure for the encapsulation of the Video Coding Layer NAL Units, and Aggregation Packets for the encapsulation of the non Video Coding Layer NALUs transmitted over TCP. Each view of the testing video sequences is transmitted over the network concurrently using UDP/IP connections to the client. In order to emulate background traffic and to stress the performance of the system, Dummynet [36] is set up between the streamer and the client. The test-bed platform is illustrated in Figure 14, while Table 5 summarizes the encoding and streaming characteristics.

Table 5: Encoding and Streaming Parameters

Encoding Parameters			
Video Sequence	Martial Art	Munich	Panel
No. of Frames	1102	1463	1000
Intra Period	8	8	8
Quantization Param.	36-30-24	36-30-24	36-30-24
Frame Rate	25	25	25
Resolution	1920x1080	1920x1080	1920x1080
Packetization Parameters			
Video Packetization	NALU < 1460 bytes		
RTP Packetization	Single NAL Unit		
Streaming Parameters			
MTU size	1500 Bytes		
Packet Loss Rate	0%	2%	4%	6%

In order to produce reliable and repeatable results, the subjective analysis of the video sequences has been conducted in a controlled testing environment. An LG W2363D 3D LCD monitor with resolution 1920x1080, aspect ratio 16:9, peak display luminance 400 cd/m2, contrast ratio of 1000:1, along with active shutter glasses is used during the experiments to

display the stereoscopic video sequences. The viewing distance for the video evaluators is set to 1m, in accordance with the shutter glasses manufacturer specifications. The display is calibrated using the Spyder 4 Elite Monitor Calibration tool. The measured environmental luminance during measurements is 200 lux, and the wall behind the screen luminance is 20 lux, as recommended by [37].

During the subjective assessment of the perceived video quality, the observers are rating the video sequences according to the Double Stimulus Continuous Quality Scale (DSCQS) method as described in [4]. According to DSCQS methodology, two consecutive presentations of original reference video (Stimulus A) and the distorted video (Stimulus B) take place with a duration of 10 seconds for each stimulus and 1 second separation in between the presentation. This procedure is repeated two times, and at the last round the observer must rate the perceived quality of both the reference and the distorted video in the scale 0 to 100, where 0 represents excellent image quality perception and 100 bad image quality perception. The rating scale is also labeled with adjective terms [bad], [poor], [fair], [good] and [excellent]. The subjective evaluation was performed by 20 experts and non-experts observers. A training session was included in order for the observers to get familiar with the rating procedure and understand how to recognize artefacts and impairments on the video sequences. Finally, the resulting scores are statistically processed and the subjects whose scores deviate from the scores of the other subjects are removed using the technique of outlier detection. The outlier detection refers to the detection and removal of scores in cases where the difference between mean subject vote and the mean vote for this test case from all other subjects exceeds 15%.

4.4 Objective and Subjective Results

The following Figures Figure 15, Figure 16 and Figure 17, illustrate the obtained objective VQM of the left and right views, the subjective MOS scores and the objectively estimated QoE of the stereoscopic videos. For better observation of the results and in order to compare both objective and subjective metrics, the MOS scores have been reversed, thus *0* MOS reflects Excellent quality, while *100* MOS Bad quality. All the values shown in the figures are the average values of several subjective experiments conducted under identical conditions.

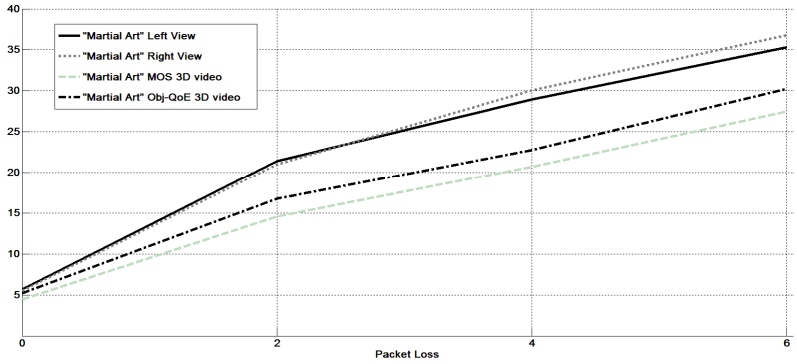


Figure 15: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model ("Martial Art" sequence)

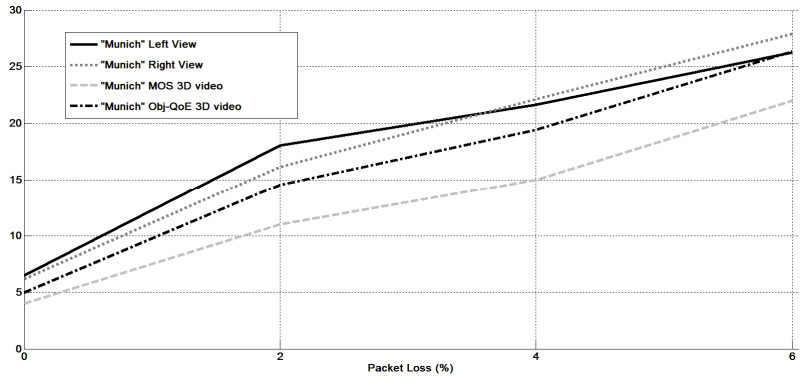


Figure 16: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model ("Munich" sequence)

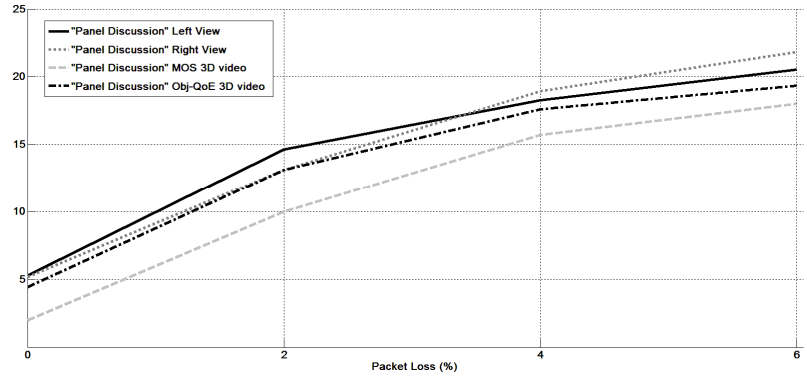


Figure 17: Comparison of VQM metric against MOS scores for stereoscopic video and the proposed Objective QoE model ("Panel Discussion" sequence)

4.5 Analysis of Subjective results

The SVC encoding of the left and the right view and their unicast transmission over identical and statistical independent channels resulted in both views to experience almost the same average distortion. Therefore, as it can be seen in the results, the VQM plots of both views are very close under all packet loss conditions. Moreover, it is shown that the video sequence of the highest motion ("Martial Art") is grater affected by packet loss, hence has the highest VQM values (i.e. lower perceived quality) compared to the other sequences under the same networking conditions.

Additionally, MOS scores of the stereoscopic video can be seen that are correlated with the VQM values of the 2D videos, which is expected as VQM can be calibrated based on the testing video sequence's particular characteristics in order to produce high quality ratings. The fact that for all three testing sequences and under all network conditions MOS ratings are better than the corresponding VQM values, indicate that the depth perception of the 3D video increases user's perception quality, resulting in higher QoE. It is important to underline the fact that MOS scores increase (i.e. the user experience worsens) almost linearly with packet loss in the case of "Martial Art" sequence, as opposed to the other two sequences.

Finally, the objective QoE model shows a very close correlation with the VQM values, as well as the MOS scores. The fact the *Obj_QoE* plot lies in between the VQM and MOS plots is in accordance to the fact the since it is based on objective measurements is more strict than MOS in estimating the quality of video user' perception. It can be seen that the proposed model is closer related to MOS than VQM due to the fact that incorporates a weight factor *A*, which reflects the human factor during the QoE estimation.

Nevertheless, the proposed model requires further study and future work has to focus towards rendering it capable of being applied to any video sequence. One of the main aspects that have been overlooked is the validation of the model with colour plus depth cues. Detailed user

profiling, more closer study of human factors including psychological and social, as well as, incorporation of environmental parameters into the proposed model is part of an ongoing research. Furthermore, networking parameters such as end-to-end delay, packet jitter, UEP overhead, the effect of P2P overlay re-organization and chunk scheduling and handoff latency that are present in the ROMEO architecture, need to be included in the proposed model. Therefore, this study is far from complete and further investigation is already underway.

5 MEASUREMENT OF RENDERING ARTEFACTS FOR MULTIVIEW VIDEO APPLICATIONS

5.1 Introduction

In this section we will discuss the methodology for evaluation of depth map quality. We will briefly discuss the system architecture, while presenting the reasoning behind adopting our approach, furthermore we will discuss the disparity-to-depth conversion required to implement the software module for the case of disparity map based rendering that is envisaged in ROMEO.

5.2 Metrics/Software for measuring the quality of depth maps for rendering

Here we will present the process implemented in this project to evaluate the quality of the depth map. In this module we will utilise the standard MPEG depth based view synthesis software (ViSBD 2.0) [38], this will be the main working block within our system. ViSBD will be used to synthesis a virtual view at the position of an already existing view. The rendering process will be performed using one (or two) coloured view(s) and its associated depth map(s). In other words a synthesised view at Camera 2 will be created form colour and depth information available at Camera 1 (or 3).

The next stage after the rendering step, the quality of the synthesised colour view has to be evaluated by comparing this view to the original at that same location, this is done through an established Objective Measure. In the current stage of the depth quality assessment module we have decided to employ PSNR as a measure of spatial quality and Temporal- Peak Signal to Perceived Noise Ratio (PSPNR) as the measure of indicating temporal quality. PSPNR measure takes in to account the just noticeable difference in luminance contrast and does not consider small pixel displacement errors caused by rendering.

Prior to running the objective measurement the issue of the disocclusions (holes) present, as an inherent problem of DIBR, in the synthesised view has to be addressed. As one of the options in ViSBD 2.0, a hole mask can be outputted in the form of a YUV sequence, this mask represents the locations of the hole positions [38]. Therefore the hole mask will be applied automatically within the depth quality module to both the virtual and original views then they will be forwarded as masked versions to the objective measurement state. The system architecture of our proposed system is illustrated in Figure 18 below.

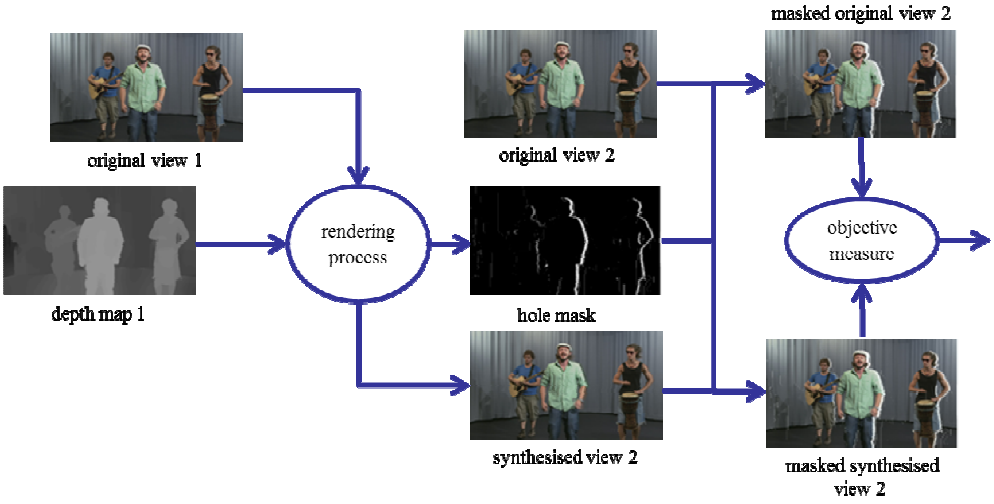


Figure 18: System architecture of Depth Map Quality measurement

The above figure demonstrates the whole process of our system, which we will refer to in the rest of this report as Depth_Map_Qual (DMQ). Although the above architecture demonstrates the use of a single colour view and its related depth map in the rendering process, this system can be used to evaluate the quality of depth map in a system where two colour views and their associated depth maps can be used in the synthesis process (i.e. views at locations 1&3). Keeping in mind that the later scenario is the more realistic, as most rendering applications are used to interpolate intermediate views rather than extrapolating views outside the main MVD baseline, it is still important to judge the contribution of individual depth maps to the overall quality, hence the need for the single depth map scenario.

An inherent problem of DIBR is the disocclusions commonly referred to as holes'. Basically holes occur because parts of the viewing area, which are occluded in the reference view, become visible in the targeted virtual view. This is demonstrated by Figure 19 below.

Figure 19: Holes caused by disoccluded regions. (a) Cause regions (b) Virtual left view of 'Ballet' sequence (white pixel are the holes)

From the above figure it is evident that even assuming a perfect depth/disparity map the holes will still occur, due to the fact that there is missing information present in the required view, which is not available in the reference view. On the other hand, most rendering algorithms compensate for this missing information and even for the holes occurring, in the case interpolating intermediated views, which are caused by conflicts occurring due to the mapping, in the rendering process, of two pixels to the same location in the target view [39]. To eliminate the missing information and compensation factors from our measurement the hole mask referred to in the previous section is applied to both original and synthesised views before implementing our objective measure.

Since we have introduced the methodology behind our quality module, it is necessary to mention briefly the step utilised for the accommodation of using disparity maps in the rendering process. Since we are using the MPEG standardised software ViSBD 2.0 for the view synthesis process it is necessary to convert the disparity maps, which might be available into suitable depth maps.

The conversion of the disparity maps available for the MUSCADE sequences in our video data set, was performed using a MATLAB code, which processed the original Disparity data files and extracted the necessary information required for the production of an 8-bit 4:2:0 YUV depth map file [40], which is required by the rendering software. In equation (1) below we find the operation carried out to convert the disparity values into depth values. It is worth mentioning that only the 8-bit disparity values within the files was extracted, since it contains the integer disparity values.

$$d = f * \frac{b}{disparities} \qquad (1)$$

So the above equation was used to convert the integer disparity values into depth values, where *d* is the depth value (Z), *f* is the focal length and *b* is the normalised baseline. Due to the fact that a disparity value of zero existed within the original disparity maps, we replaced these zero values with a value of one to eliminate the division by zero issue. Then to identify the near and far depth plane, we calculated the maximum and minimum of the outcome depth values (d) form equation (1). These maximum and minimum depth values were set as Z_{far} and Z_{near} respectively (far and near clipping planes), then those values where applied to equation (2) below, in order to extract the corresponding *V(x,y)*, where (x,y) indicate the exact pixel location within the final depth map.

$$V(x,y) = 255 * \frac{\frac{1}{Z(x,y)} - \frac{1}{Z_{far}}}{\frac{1}{Z_{near}} - \frac{1}{Z_{far}}} \qquad (2)$$

Finally, after obtaining the values form the above equation the 8-bit depth values were written into a 4:2:0 YUV file for each single frame and then the frames were merged into one single YUV sequence in order for it to be utilised for the purpose of rendering the virtual views.

5.3 Objective Results and Discussion

In this section we will present the results of the DMQ module being run on the MUSCADE video sequences but first we will explain the principle behind the testing process. After the disparity -to-depth conversion the resulting depth maps were encoded at different quantisation levels (QP), in order to introduce quantisation errors to the available depth/disparity maps in order to test the quality of depth maps present. The encoded structure used was an I & P frame structure, with the intra-frame period set to 8. Three QP levels were chosen randomly (QPs chosen were: 22, 32 and 42) and all results decoded depth/disparity maps, together with the original one was used for the rendering purposes.

The naming notation used for the sequences rendered will be referred to as: Orig, IP2222, IP3232 and IP4242 in the rest of this report. The view to be synthesised was chosen as view 3 and the rendering process was performed using views 2 and 4 (with their associated depth maps) separately and both views were used to synthesis view 3 together as well. The results obtained by DMQ are illustrated in Figure 20 below.

From the Figure 20, it is evident that the different compression levels, applied to the depth maps, have some impact on the quality of the rendered view, although it seems that a up to a certain level of compression the degradation in view quality is less than that of the highest compression level. As well the loss of quality is demonstrated in both the PSNR and PSPNR measurements, which indicates that the quality of the depth map has an impact on both spatial and temporal quality of the synthesised view.

It is also worth mentioning that the contribution of each depth map, when employed individually to the quality of the rendered view is affected by the distance to that view (i.e. the distance between view 4 and view 3 is much larger than the distance between view 3 and view 2), in other words the ability of the depth map to shift the colour pixels of the reference view onto the desired location is affected by the distance between the reference and virtual views (this issue will be touched upon in the future work section of this report).

However, when taking a closer look we can notice that the effect of the distance on the rendering varies between the different sequences tested above. We notice that the Band sequence performs a lot better than the BMX and Musicians sequences, which have quite a similar performance level, despite the fact they all have a near identical physical setup (apart from a very small variation in the baseline). This can be attributed to the quality of the original depth maps of these sequences, hence raising the question about the depth/disparity map creation procedure for both of the BMX and Musicians sequences (in fact the disparity information for both those sequences was processed at a different phase of the MUSCADE project to that of the Band sequence).

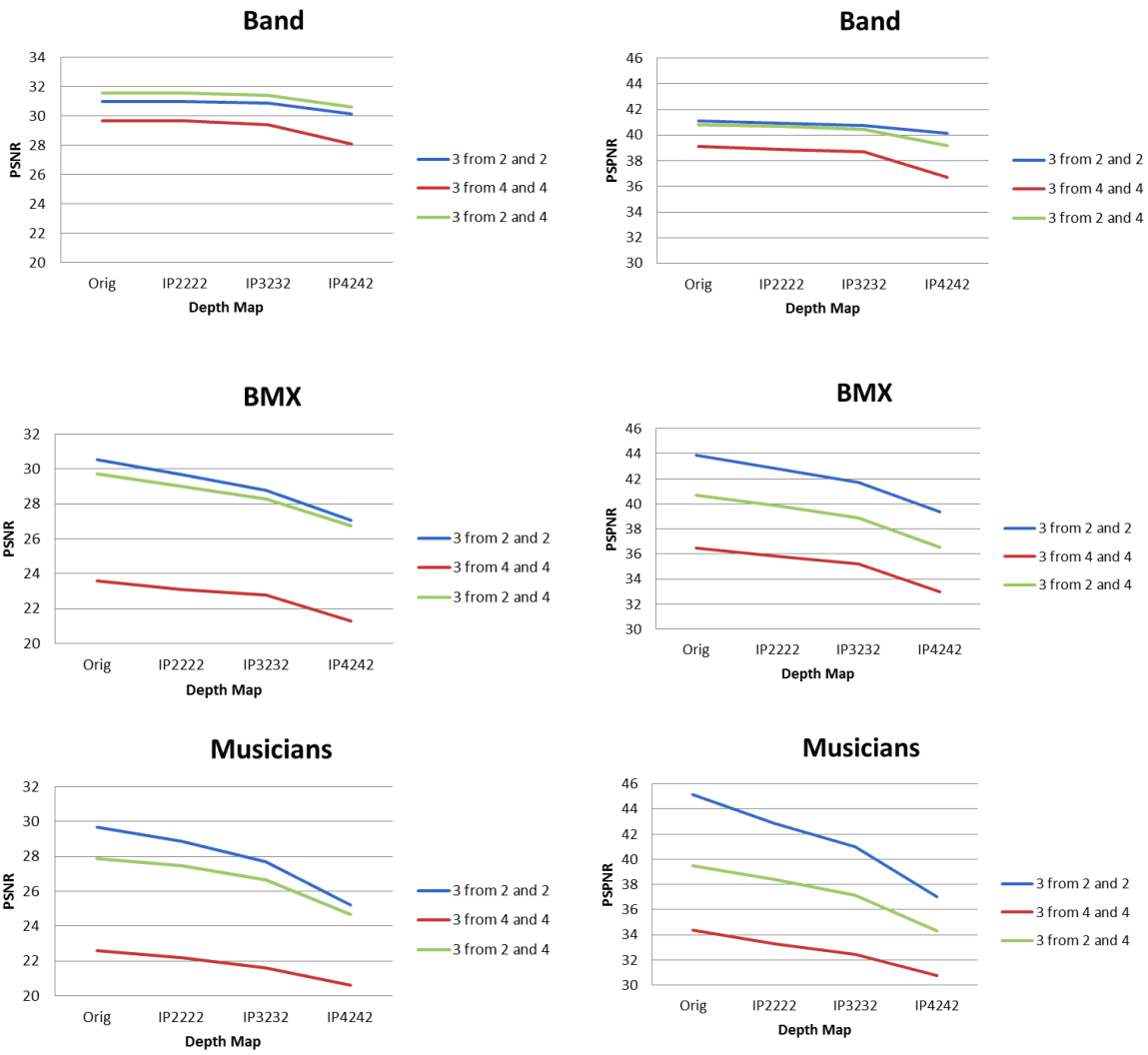


Figure 20: DMQ Results for rendering with single disparity/depth map

Finally to have a better idea about the comparison between those three sequences, Figure 21 below illustrates the performance of those depth maps (interpolating scenario) in comparison to each other. The argument for the case of better depth quality in the Band sequence is confirmed by the PSNR readings and although the temporal PSPNR measure does not offer much of a difference between those depth maps, when the original maps are used, the Band results present less degradation of quality at the compression levels, while the other two sequences rendering temporal quality deteriorates in larger manner.

This issue of different depth map quality for the various sequences employed and its creation techniques will be mentioned, together with a suggested plan to tackle this problem in the future work section of this confirmation report.

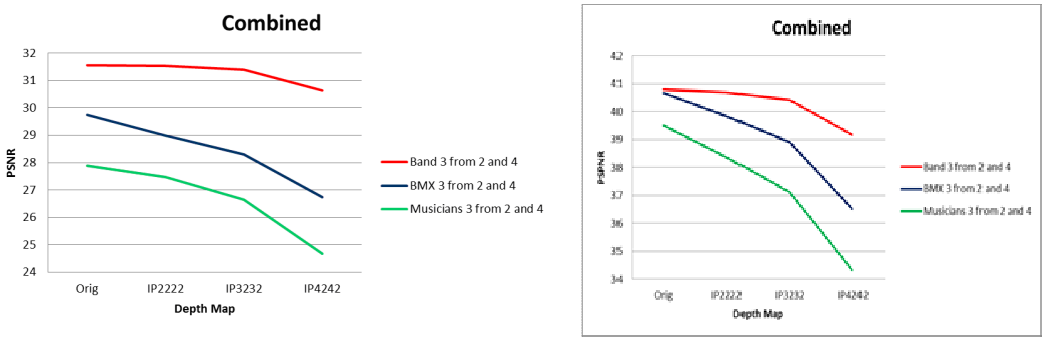


Figure 21: DMQ Results for double disparity/depth map

5.4 Subjective Assessment of Rendering Artefacts

One of the most challenging issues within this research is the issue of carrying out a valid subjective experiment, keeping in mind the inherent problem of DIBR, which is disocclusions and as put by [41] that 'synthesised views contain specific artefacts around the disoccluded areas'.

Especially in the single depth/disparity map scenario, such view synthesis artefacts, which are visually significant, may mask the quality of the distortions produced by the difference in the quality of the depth maps used in rendering. This was confirmed in a dry run of a subjective assessment, in which the same views rendered for the DMQ purpose, plus the original targeted view, were shown in random order, in single stimulus subjective assessment (a continuous scaling method was used). The results of this mini experiment were inconclusive and confirmed the fact that the presence of such view synthesis artefacts will mask other distortions related to depth map quality, even if it is picked up by an objective measure.

In the case of the use of two views and their related depth maps (interpolation) to render an intermediate view the view synthesis artefacts are much less visible and the depth map quality effect on the overall quality of the rendered view is much more visible. However a subjective assessment for the interpolation of intermediate case was not run, pending a solution to the extrapolation case and the addition of extra testing materials.

5.5 Future work on depth/disparity map quality evaluation

In this section we described the software developed to measure the quality of depth/disparity maps in terms of its rendering ability. However, the experiments performed so far are based only on various compression levels of depth maps. Therefore, to be able to measure quality of different depth maps generated using different estimation techniques and processed with different post-processing schemes, we envisage developing metric considering different view synthesis artefacts. The main difficulty of a view interpolation is how to cope with occlusion errors, and since the present approach disregards the occlusion effects, in the future we will also take in to consideration the hole filling approach as well as the amount of disocclusions. In the future we would run the subjective experiments where a particular view of a video is rendered with different depth maps to isolate subjective perception of view synthesis artefacts.

6 MEASUREMENT QOE OF SPATIAL AUDIO

This section introduces the concepts and key attributes of the Quality of Experience (QoE) of spatial audio, and describes findings from an initial set of experiments. Furthermore, to aid the selection of audio codecs to compress spatial audio in WP4, this section also presents a comparison of audio codec performance with the aid of an objective perceptual metric.

6.1 Introduction

The audio QoE relates to all perceived aspects of audio in the contents delivered to the end users. In particular, when multichannel spatial audio is used as the media, the spatial attributes of perceived sound become meaningful, as well as the timbral attributes which have already been known as the acoustical characteristics related to QoE. In addition, when the audio accompanies video, the correlation between the audio and video contents can also affect QoE. Modelling and prediction of QoE involves identifying the physical characteristics of sound that are related to the perceived audio quality and finding specific relationship between them through subjective tests to be able to use they physical parameters to describe the QoE as perceived by actual users. The following subsections describe the three attributes that will be used in combinations to describe the audio QoE in the project, and introduce the physical parameters of sound that can be used to predict these.

6.1.1 Spatial audio QoE attributes

An extensive study previously conducted on the perceptual attributes of spatial audio [42] reviewed related previous works extensively and included detailed explanations of the spatial audio QoE attributes, except for audio-video correlation which is a relatively new concept specific to the video-accompanying audio media.

6.1.1.1 Timbral Fidelity

Before the introduction of multichannel (above stereo) spatial audio, an attributed named Basic Audio Quality (BAQ) represented an attribute that corresponds to all the perceivable differences between a coded (altered) audio signal and a reference signal. This is mostly related to the frequency contents of the audio signal and does not take into account the spatial attributes. The introduction of the multichannel configuration made this attribute represented better with the term Timbral Fidelity (TF). An objective model exists for the measurement of the basic audio quality as defined in Rec. ITU-R BS.1387-1 method for objective measurements of perceived audio quality (known as PEAQ) [43]. Previous related studies have found out that BAQ is highly correlated with TF [44].

6.1.1.2 Spatial Fidelity

Spatial Fidelity is defined for multichannel reproduction configurations, when there are sounds coming from the rear. [42] divided it into two, particularly for the standard 5.1-channel configuration: Front Spatial Fidelity (FSF) and Surround Spatial Fidelity (SSF). FSF is related to the spatial aspects of the sound perceived in the frontal plane defined by an arc between -30 and +30 degrees. SSF is related to the spatial aspects of the sounds perceived outside of this arc. This division is made considering the facts that human sound source localisation is known to be more sensitive in the frontal plane [45], and that the frontal plane is also where the visual cues are available.

6.1.1.3 Audio-Video Correlation

This attribute refers to the spatial correlation between the audio and video, when the spatial audio is presented with the video. The spatial mismatch between the sound objects seen in the video and the corresponding audio may affect the overall QoE. This is valid only when the video is present, and therefore has not been investigated as much as the other QoE attributes introduced above.

6.1.2 Acoustical parameters for QoE prediction

Previous studies have found the low-level parameters that can be calculated from the given audio signal to be matched with the subjective grading scores of TF and SF.

6.1.2.1 Spectral coherence

Coherence spectrum describes the correlation between the reference (original) and test audio signals at certain frequency indices. It is given as follows:

$$Coherence = \frac{\|P_{rt}(f)\|^2}{P_{tt}(f)P_{rr}(f)} \quad (1)$$

where *P(f)* is the power density and the subscripts denote the power density of the test sequence, the reference sequence or the cross density of reference and test sequences.

6.1.2.2 Spectral roll off

Spectral Roll off (SR) is given by the frequency index where 95% of the total spectral magnitude is obtained. Using the STFT to calculate the magnitude spectra for each frame as in (1), SR is the average value of the individual Spectral Roll-off values for the frames which is calculated by taking the minimum SR_j satisfying (2).

$$\sum_{n=1}^{SR_j} M_j[n] \geq \sum_{n=1}^{N} M_j[n] \times 0.95 \quad (2)$$

where $M_j[n]$ is the magnitude of the STFT of the frame *j* over blocks and *n* is the frequency index. SR is then found by averaging the values across all the frames. The Spectral Rolloff values need to be rescaled prior to use in the final model. The rescaling is done by calculating the values for the original (hidden reference) signal and the 3.5-kHz low-pass filtered version (anchor). The rescaled roll off is then given by:

$$SR_{rescaled} = x|SR| + y \quad (3)$$

where

$$x = \frac{80}{SR_{ref} - SR_{anch}} \quad (4)$$

and

$$y = \frac{20SR_{ref} - 100SR_{anch}}{SR_{ref} - SR_{anch}} \quad (5)$$

6.1.2.3 Interaural cross correlation (IACC) at 0 degree head rotation

This is known to be related to “spaciousness” or “apparent source width” of sound, and is calculated as:

$$IACC_{bb0}(\tau) = \frac{\int_0^T P_L(t)P_R(t+\tau)dt}{\sqrt{\int_0^T P_L^2(t)dt \int_0^T P_R^2(t)dt}} \quad (6)$$

where L and R denote the left and right channels of binaural recording. The binaural signals can be obtained by passing the loudspeaker signals through the Head-Related Transfer Functions (HRTFs) corresponding to the loudspeaker directions. τ is varied within the range $\pm 1ms$±1ms. For each frame the maximum value of these calculations is used and as the $IACC_0$ IACC$_0$ value, the average is taken across all frames. This value can be rescaled in the same way the Rolloff is rescaled, this time to a range of 1 and 0 using the same formula given

in (3), except for the fact that the IACC value for the mono anchor is 1. The rescaled version of the broadband IACC at 0 degrees then becomes

$$IACC_{bb0,rescaled} = \frac{|IACC_{bb0}| - IACC_{bb0,ref}}{1 - -IACC_{bb0,ref}} \quad (7)$$

6.1.2.4 Maximum of IACC at 0, 90, 120, 150 and 180 degree head rotation

To calculate these parameters, the signal first passes through three octave filters with centre frequencies at 500, 1000 and 2000 Hz. Then the same procedure as described above is applied to calculate the rescaled low-frequency band IACCs, and the maximum of these three is taken for each frame.

6.1.2.5 Back-to-front energy ratio

This parameter is known to be related to the feeling of envelopment, and is calculated as

$$BFR = \frac{E_{back}}{E_{front}} \quad (8)$$

where E denotes the sum of RMS levels in the front or rear loudspeakers.

6.2 Subjective Test for Spatial Audio QoE Prediction

This section introduces a subjective test conducted at US to investigate the combined effects of the acoustical parameters described in the previous section on the perceived spatial audio QoE, and to find a prediction model.

6.2.1 Experiment design

The relationship between the QoE and the introduced attributes TF, FSF and SSF have already been found in [42]. Therefore, in this experiment the acoustical parameters known to predict TF, FSF and SSF were controlled, with the addition of an Audio Video Correlation (AVC) related parameter. It was considered reasonable at the current stage to use the angular mismatch between the audio and the objects as the parameter indicating AVC, as the most prominent factor when watching video contents accompanied by multichannel spatial audio.

Firstly, in order to introduce AVC degradations, a visual cue of the audio objects was created to be presented. Then the auditory scene was created such that there is a varying angular mismatch between the visual cue and the perceived audio. The visual cue was prepared using Google SketchUp, to show the subjects the layout of the auditory scene originally intended without any angular deviation. The basic stimuli were created using an acoustic simulation and auralisation software Odeon. It was assumed that three different musicians – a soprano singer, a violinist and a violist were playing on the stage, 4.76 metres apart from each other. The listener position was set at the front centre of the audience seating area, 5.3 metres away from the central sound source: the soprano singer. Short excerpts from anechoic recordings of the individual musicians playing a classical music piece (an aria of *Donna Elvira* from the opera *Don Giovanni* by Mozart) were used in the auralisation. The resultant sound field, containing the direct sounds, the reflections and the reverberation, was simulated suitable for 5.1-channel surround sound reproduction. The initial stimulus was created assuming that the listener was facing the central sound source.

Secondly, the degradation methods for the other attributes – BAQ (non-multichannel), FSF and SSF – can be found in the literature [42]. Bandwidth limitation through low-pass filtering is a general method to vary perceived TF. FSF and SSF can be varied through different down-mixing algorithms, assuming a standard 5.1-channel reproduction system. The variations introduced to create degradations in the KPIs were determined as follows:

- 5 angular deviations (5, 15 and 25 degrees to the left, and 10 and 20 degrees to the right)

- 3 bandwidth limitation algorithms, using filters with various cut-off frequencies applied at each channel (13th order IIR Chebyshev Type 1 filter with passband ripple equal to 0.1dB)
    - All 12000Hz
    - Hybrid A (Left, Right – 20kHz; Centre – 10kHz; Left Surround, Right Surround – 5kHz)
    - Hybrid F (L, R - 10kHz; C - 13kHz; LS, RS - 3.5kHz)
- 2 down-mixing algorithms applied to the 5 channels (initially 3 (front)/2 (rear) setting)
    - 3/0 format (L' = L + LS×0.7071, R'= R + RS×0.7071, C' = C)
    - 2/0 format (L' = L + C×0.7071 + LS×0.7071, R' = R + C×0.7071 + RS×0.7071).

For each type of degradations there is an unprocessed original. This leads to the total number of combinations of (5+1)×(3+1)×(2+1) = 72.

The methodology introduced for the subjective test was based on MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) defined by ITU-R recommendation BS.1534-1. Following the guideline, a reference and two anchors were introduced additionally. The anchors were the low-pass filtered (at 3.5kHz) version of the 25 degree-shifted stimulus (no down mixing applied), and a down mixed version of the 25 degree-shifted stimulus (to mono, 1/0 format). For the listening test interface, 12 pages consisting of 9 stimuli have been designed. On each page were 6 processed stimuli, the reference and the two hidden anchors. The evaluation was divided into 2 sessions, in each of which the subjects were asked to listen to a total of 6 pages of stimuli. *Figure 22* shows the user interface for the subjects to grade the perceived QoE.

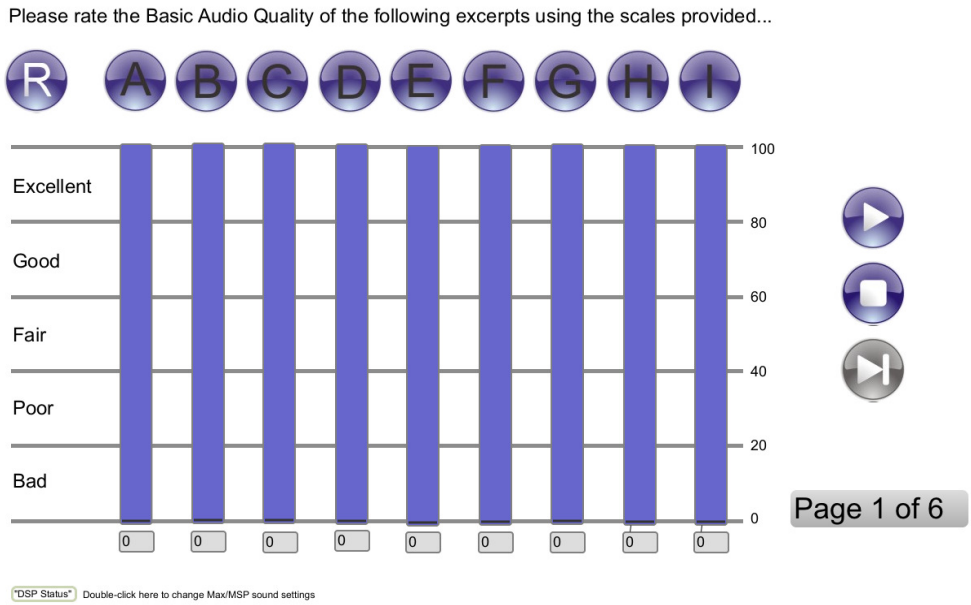


Figure 22: User interface for subjective tests on Audio QoE. On each page there are 6 processed stimuli, 1 hidden reference and 2 anchors.

6.2.2 Test environment and procedure

The subjective test was conducted in a studio built in the I-Lab, University of Surrey with a dimension of 4.1m × 2.5m × 2.1m. The environment contained an acoustically transparent screen of size 2.40m × 1.35m, on which the visual cue was projected. 16 subjects with normal hearing participated in the listening test. Each subject was seated on a chair to face the middle of the screen 2 m away from it. The loudspeakers were placed around the subject for standard

5.1-channel reproduction, with 5 channels over a circle with the radius of 1.8m and with the subwoofer at the front right corner. The user interface was provided on a separate screen of a laptop. The subjects were firstly given an instruction describing what kind of degradations were introduced in the test, and then asked to grade the overall perceived audio quality (QoE). A short familiarisation session was given prior to each test session, to help the subjects know what to expect during the actual evaluation.

6.2.3 Listening test results and findings

The answers from the subjects were collected and analysed. Firstly, the effects of AVC were examined. Then an attempt was made to find a prediction model to find the relationship between the parameters controlled and the QoE scores through linear regression.

6.2.3.1 Effects of AVC on QoE

Firstly, the result of correlation analysis, between the QoE score and the magnitude of angular deviation of auditory scene from the presented video, is listed in Table 6. The total number of observed pairs of values is 1728. It is seen that the QoE and the angular deviation is statistically significantly correlated with a coefficient of -0.557, which indicates that the QoE score will become lower as the angular deviation increases. This tendency is shown in *Figure 23*. In particular, the perceived QoE degradation is statistically significant and large as the angular mismatch is firstly introduced at as small as 5 degrees. For the angular mismatch from 5 to 20 degrees, the perceived QoE seems to decrease, although the tendency is not statistically significant. As the angular mismatch increases from 20 to 25 degrees, the QoE degradation is again statistically significant and the most noticeable. This implies that the AVC degradation, if unavoidable, should be kept such that the angular mismatch between audio and video is less than 20 degrees.

Table 6: Result of correlation analysis using SPSS between the QoE score and the magnitude of angular deviation of auditory scene from the presented video.

		Ang_Dev_abs	QoE_Score
Ang_Dev_abs	Pearson Correlation	1	-.557**
	Sig. (2-tailed)		.000
	N	1728	1728
QoE_Score	Pearson Correlation	-.557**	1
	Sig. (2-tailed)	.000	
	N	1728	1728

**. Correlation is significant at the 0.01 level (2-tailed).

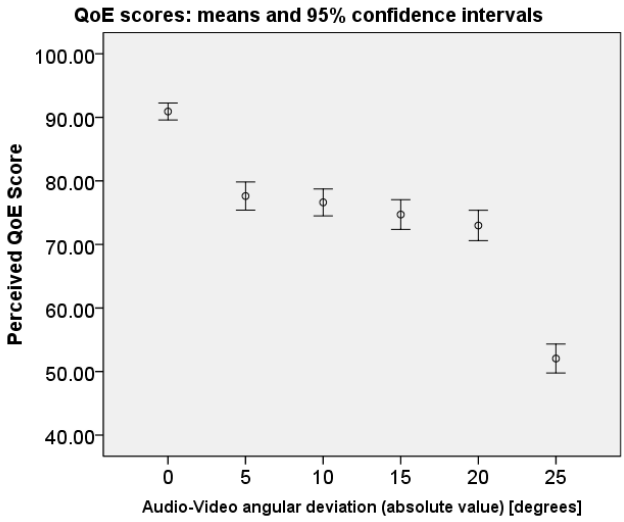


Figure 23: Means and 95% confidence intervals of QoE scores for different angular deviation of auditory scene from the video.

6.2.3.2 QoE prediction model derivation

In order to derive the QoE prediction model, regression analysis was conducted using the parameters introduced in Section 6.1.2 as the independent variables, and the QoE score as the dependent variable. The independent variables and are listed below with brief descriptions:

- Broadband IACC (Interaural Cross-Correlation Coefficient) for head orientations at 0° and 90°: **Ibb0, Ibb90**
- Centroid of spectral coherence between test and reference signals: **COH**
- Maximum octave band IACC: **Ixx** (at various head orientations)
- Back-to-front energy ratio: **BFR**
- Spectral roll off (related to upper cut-off frequency of signal): **R**
- Spectral centroid (centre of gravity of spectrum): **C**
- Angular mismatch between audio and video: **AVCang**
- Normalized spectral roll off difference between reference and test stimulus: **Rdif**

Ridge regression was conducted in SPSS in order to predict the QoE score as a linear function of these variables as follows:

QoE_spatial_audio = f(Ibb0, Ibb90, COH, Ixx,…, BFR, R, C, Rdif, AVCang).

Ridge regression is known to be effective when regression analysis is conducted using independent variables that might potentially have linear inter-relationship [46]. The ridge regression was iterated, during which independent variables that have strong linear inter-relationship were discarded from modelling. The variables that showed ignorable importance in the prediction accuracy were also discarded through the iterations. During the iterations the correlation and standard error values were monitored such that the prediction accuracy was not degraded by discarding the independent variables. Table 7 shows the result of the ridge regression from SPSS, after the iterations. B denotes the coefficient of each variable in the prediction model, SE represents the standard error, and Beta indicates the relative importance of the corresponding independent variable (excluding constant). The correlation between the predicted QoE and the subjective QoE scores was found to be high at 0.89445, with the standard error of 9.0674.

Table 7: Result of ridge regression using SPSS, after iterations to reduce the number of independent variables.

	B	SE(B)	Beta	B/SE(B)
lbb0	-5.68904	1.679845	-0.12276	-3.38665
lbb90	-18.6713	3.308319	-0.22244	-5.64375
COH	0.003572	0.000236	0.523401	15.11125
l0	-1.05482	0.704304	-0.05612	-1.49768
BFR	-7.11362	1.41538	-0.20669	-5.02595
Rdif	10.35965	3.173443	0.11465	3.264483
AVCang	-0.75263	0.073681	-0.39751	-10.2148
Constant	60.77483	2.777324	0	21.88252

In Table 7, B denotes the coefficient of each variable in the prediction model, SE represents the standard error, and Beta indicates the relative importance of the corresponding independent variable (excluding constant)

Consequently, from the subjective test results the QoE score can be predicted using the following expression:

QoE_spatial_audio = -5.68904·lbb0 – 18.6713·lbb90 + 0.003572·COH – 1.05482·l0 – 7.11362·BFR + 10.35965·Rdif – 0.75263·AVCang + 60.77483.

Figure 24 shows the scatter plot drawn to check the distribution of the subjective QoE scores against the predicted values.

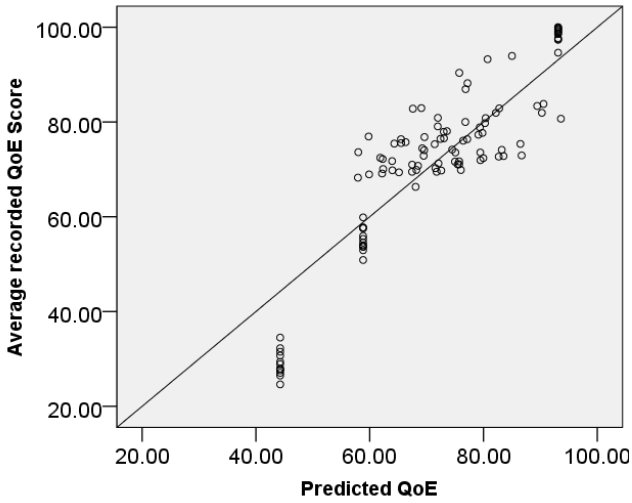


Figure 24: Scatter plot of the subjective QoE scores averaged across the subjects, against the predicted QoE values using the results of ridge regression.

It is seen that the distribution follows the y=x line closely, confirming the high correlation between the subjective answers and the predicted QoE scores.

6.3 Comparison with SoA in audio QoE and new avenues for prediction model refinement

As described previously, this experiment was conducted as a preliminary step to investigate the relationship between the spatial audio QoE and the acoustical parameters known to predict its key attributes, especially including AVC as a new factor. This approach was applied along with the initial model to the MUSCADE project towards the overall QoE prediction. It is worth noting that the audio quality prediction model previously developed before the introduction of AVC had not taken the presence of video into consideration, and had assumed 5.1 channel-based spatial audio. In this case, FSF and SSF are mainly related to spatial processing such as down mixing in the frontal and rear areas, whereas AVC is controlled by the angular shifting of the whole auditory scene.

On the other hand, a slightly different approach had been taken for the audio QoE modelling for the DIOMEDES project, although the categorisation of the QoE attributes was similar to what was introduced in Section 6.1.1. This was because the reproduction system was object-based WFS (Wave Field Synthesis) system which has a different loudspeaker layout, making the previously developed concepts of Spatial Fidelities irrelevant. However, it was noted in this case that AVC in its original concept could have overlapping meanings with SFs (for instance, AVC as the misalignment between the audio and video objects, and FSF as the spatial distortion from the reference in the frontal arc).

In addition, the control of TF was also different in the two projects. Whereas in MUSCADE the low level spectral parameters had been used, the DIOMEDES project had firstly investigated the effect of coding bit rate on the (timbral) audio quality and simply used it as the variable. Based on these findings from investigations in previous projects, research questions and new guidelines have been set up for the refinement of the prediction model through further experiments.

- What is the hierarchy between SF and AVC when video is present?

  More concrete hierarchy between these two attributes would help in the development of a more comprehensive prediction model with simple and clearly distinguished parameters. The presence of video now needs to be clearly considered, as well as the spatial processing possible in the audiovisual media delivery (for instance, the possibility of separate down mixing in the frontal/rear areas, or the validity of AVC behind the listener).

- Can fewer parameters be used for the QoE prediction?

  This is the question to be finally answered through the investigations in this project. Especially, more specific relationship between TF and the encoding quality would help to simplify the physical parameters related to TF, and thus to simplify the overall prediction model.

6.4 Comparison of audio codec performance for spatial audio compression

The purpose of this section is to compare the performance of the audio codecs to encode spatial audio.

Earlier studies have been reported such as in [47][48][49] showing the performance of the state-of-the-art audio codecs where each of them operating at its typical bitrates. As can be seen in Figure 25, MPEG Surround in combination with HE-AAC achieved higher perceptual quality than HE-AAC multichannel at low bitrates approximately between 64 and 128 kb/s when encoding 5.1 audio signals. However, at a bit rate of 160 kb/s, both audio codecs seem to be competitive. Moreover, the AAC multichannel audio codec is the best audio codec, in terms of the quality of the reconstructed audio signals, although it is operating at a higher bit rate that is 320 kb/s.

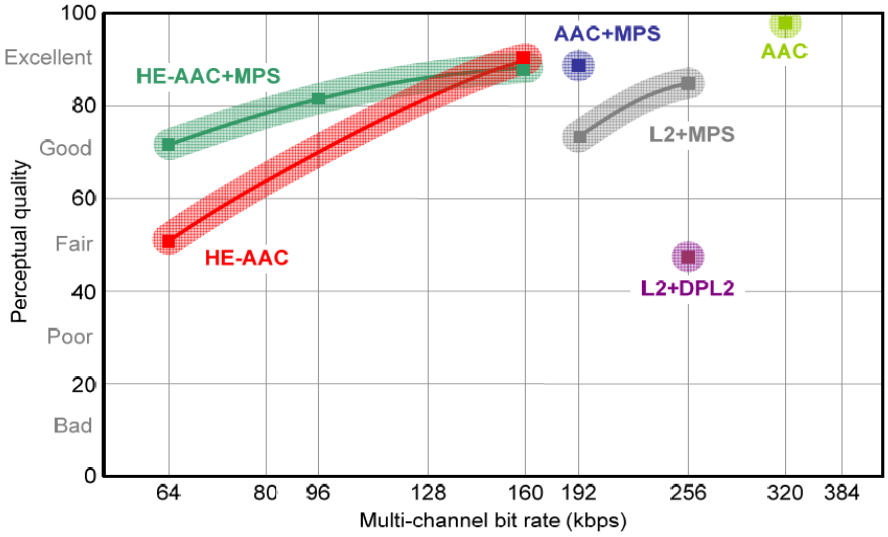


Figure 25: Performance of several audio codecs at various bitrates[47]

It is also shown in Figure 25 that MPEG Surround in combination with AAC operating at 192 kb/s is competitive to HE-AAC multichannel as well as with MPEG Surround in combination with HE-AAC. It is therefore interesting to investigate the performance of MPEG Surround in combination with AAC at higher bitrates since this is not reported yet.

The results of our experiments using ITU-R BS.1387-1 [50] for this investigation is given in Figure 26 where the objective difference grade (ODG) having five grades, that are: 0 (imperceptible), -1 (perceptible but not annoying), -2 (slightly annoying), -3 (annoying), and -4 (very annoying), are used. As can be seen MPEG Surround can achieve high performance, in terms of ODG (objective difference grade) score at higher bitrates. However, at bitrates greater than 480kb/s, AAC multichannel outperforms MPEG Surround. Based on this investigation, AAC multichannel will be adopted as the basic scheme for spatial audio coding. We plan to verify these findings with a subjective assessment.

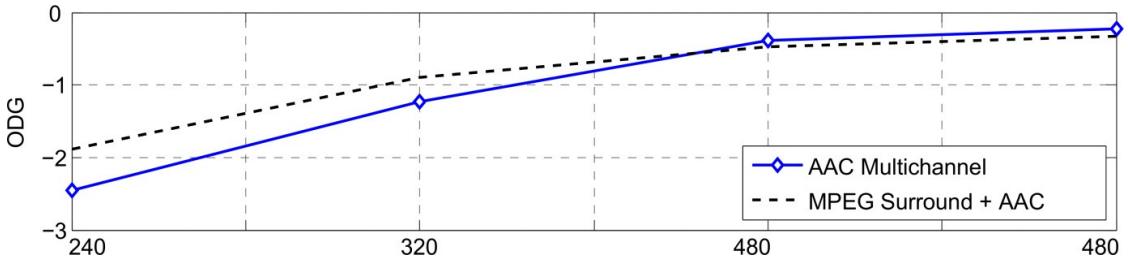


Figure 26: Performance of AAC multichannel and MPEG Surround in terms of ODG score

6.5 Future work related to audio QoE

Specific investigation will be conducted to find the relationship and hierarchy between SF and AVC. A subjective test will be designed for this purpose. The user interface needs to be enhanced so that the video can be presented instead of still images. Comparison of the predicted TF and the audio encoding quality parameter will also be conducted to help the simplification of parameters.

The initial investigation introduced in Section 6.2 used synthesised material as the stimuli (auralised audio objects). Therefore, it is desirable to conduct the validation experiment with

more general audiovisual contents. The contents captured for the project can be good examples, since the audio can be flexibly prepared in object-based configuration or in 5.1-channel configuration.

7 VISUAL ATTENTION MODELLING FOR 3D VIDEO

7.1 Introduction

Visual Attention Modeling is a process that tries to guess which elements in a visual scene will catch the viewer's attention. The capability of human visual system to notice visual saliency is tremendously fast and trustworthy. On the other hand, computational modeling of this basic intelligent behavior still remains a challenge.

In ROMEO project we have two objectives that we address with Visual Attention Modeling. First one is the improvement of video encoding and the second one content format adaptation. Both rely on the ability to detect elements in a visual scene that attract the viewer's attention. The following chapter summarizes our results of investigation and presents the concepts we are aiming to implement within the project.

7.2 Saliency Detection

Based on different methods that are summarized above we present different approaches that we intend to implement in ROMEO.

7.2.1 Visual salience based on the spatial information

Here our objective is to extract the most plausible saliency map based on the intra image information.

This approach is based on the work of Christof Koch and Shimon Ullman [27]. It takes into account factors called bottom-up, due to their low level complexity. They include three features based on the difference in term of color, intensity and orientation. They are the most attractive elements for the human brain. Figure 27 presents three images with a specific point of interest based on these features.

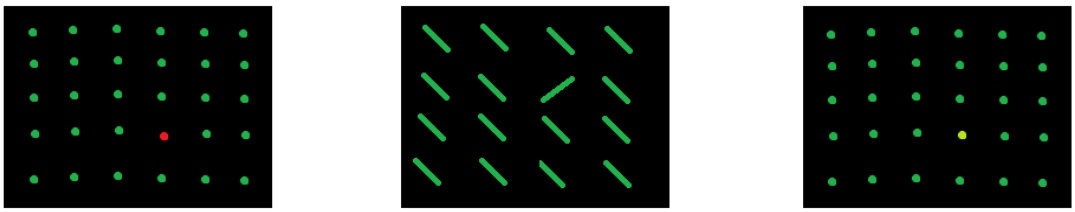


Figure 27: Most attractive features for human brain. From left to right: color, orientation and intensity.

In this analysis, it is proposed to extract maps for each feature and to apply a mechanism of Winner Take All (WTA) in order to highlight the most probable area of interest. Figure 28 is illustrating this approach.

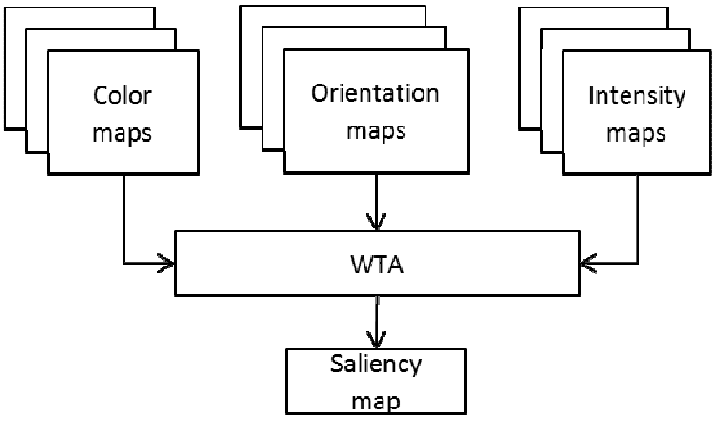


Figure 28: Overview of suggested approach.

Our work in term of innovation is divided into two steps:

- Find how the extract the main part of the information for each feature.
- Find the most relevant criterions to integrate in the Winner take all blocks.

7.2.2 Temporal Image Signature

We propose to explore the combination of Image Signature approach (IS) [28] with the Temporal Spectral Residual (TSR) [17] in order to determine the visual saliency. Utilizing the information of DCT for saliency detection is one of the main reasons of choosing this approach. The additional benefit of the DCT is its efficiency. As the size of the image increases, the FFT becomes increasingly complex at a much more rapid rate.

The suggested approach 'Temporal Image Signature' (TIS) will be a combination of IS and TSR approach. It aims at locating distinct movement. This will result in the form of salient region extraction in XT plane and YT plane. At first a number of frames will be sliced into the XT and YT planes. By doing this there is a better demonstration of the horizontal-time and vertical-time axes which also contain movement information. Then the IS approach is applied individually on all planes. This detects saliency in movements rather than in image formation. By extracting just salient movement the camera movement will remain disregarded in this step. Finally an accumulation is done after transforming back to XY domain. Figure 29 gives an overview of the approach.

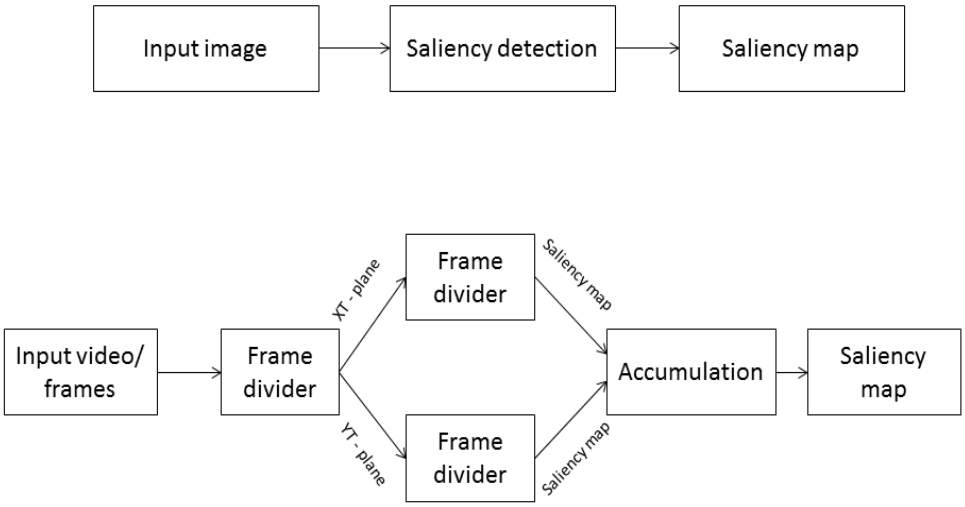


Figure 29: Overview of Temporal Image Signature. Upper diagram: Saliency detection applied on a single image. Lower diagram: Saliency detection applied on multiple frames.

Given a video clip with size m * n * t where m*n is the image size and t is the number of frames being processed, TIS can be written as follows.

$$MapXT_m = sign(DCT(I_{XT_m}))$$

$$MapXT_m \xrightarrow{transform} vMapXY$$

MapXT represents the horizontal-time map. I_{XT} is the slice or the deformed version of the images in horizontal axis. *Sign (DCT(I))* is the IS process applied on the image. The same is done on the vertical-time axis resulting in *hMapXY*. Afterwards saliency map in a temporal domain can be represented as

$$salMap(t) = hMapXY(t) + vMapXY(t)$$

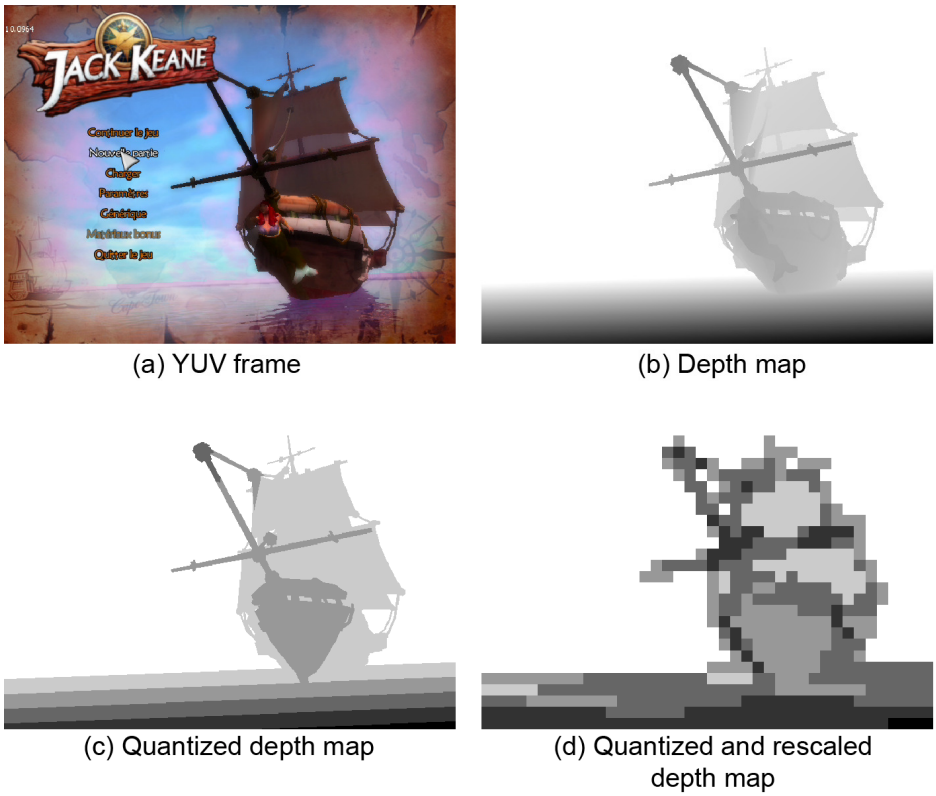
Where *t* is the number of frames being processed.

7.2.3 Depth map based saliency

In many practical cases, the objects that belong to the scene's foreground are subject to a higher visual attention from the user and are more relevant in terms of QoE. From this assumption, we can consider that the multiview depth map contains the main information when defining regions of interest. In ROMEO, the depth maps are extracted during the post-acquisition step and transmitted with the corresponding view to the encoder. So, to improve the QoE we will process this depth images in order to obtain a macroblock based partitioning of each frame as described in Figure 30.

First of all, the depth values are quantized over a small number of integer values (6 values in Figure 30 (c)). The integer value of each pixel represents its depth level. The lower values (black) correspond to the foreground and the higher values (white) to the background. As this ROI partitioning is used next to perform the bit rate adaptation during the video encoding stage, the next processing step consists in downscaling the quantized depth map values Figure 30 (d). This downscaling is basically obtained by averaging the depth values for each macroblock and by dividing this mean value by the number of different depth levels in the macroblock. After this last processing, each macroblock is represented by an integer value which is supposed to indicate its level of importance. The lower levels correspond to the higher interest for the user and this value will be used as a quantization factor offset during the video encoding stage. This last operation on the depth map allows a better discrimination of

homogeneous areas from regions with high frequencies (object borders). For instance in the given example, some object borders on the boat are almost in the background, however thanks to the division by the number of different levels, they will be better taken into account during the video encoding.



(a) YUV frame (b) Depth map

(c) Quantized depth map (d) Quantized and rescaled depth map

Figure 30: Steps from captured depth map (b) to MB based ROI partitioning (d)

7.2.4 Feature weighting

The approach “Main Subject Detection” [26] was found to be a good basis for a basic rating concept. Since we will have to handle different feature maps a rating mechanism is required that decides how to weight those maps for the final salient map. Vu and Chandler argue that a feature that provides good results on one image may be detrimental when applied on another one. They suggest using cluster density as weighting criterion. All maps are multiplied with a weighting factor w_f before they are summed up to the final saliency map. The rating mechanism has already been implemented with quite good results on single images. To adapt the approach to video we aim to extend it with both object tracking and motion estimation in a first step. Tracking for example can be used to improve the algorithms first guess of the object location. This is done by providing the tracking data for that object.

Furthermore we add some feature maps that proved to be reliable (see descriptions above). Our investigations will show if this additional input improve robustness of the final algorithm. We believe that the rating process will be a key point to quality results. Thus it will be analyzed if it can be improved. Figure 31 gives an overview of the proposed rating system.

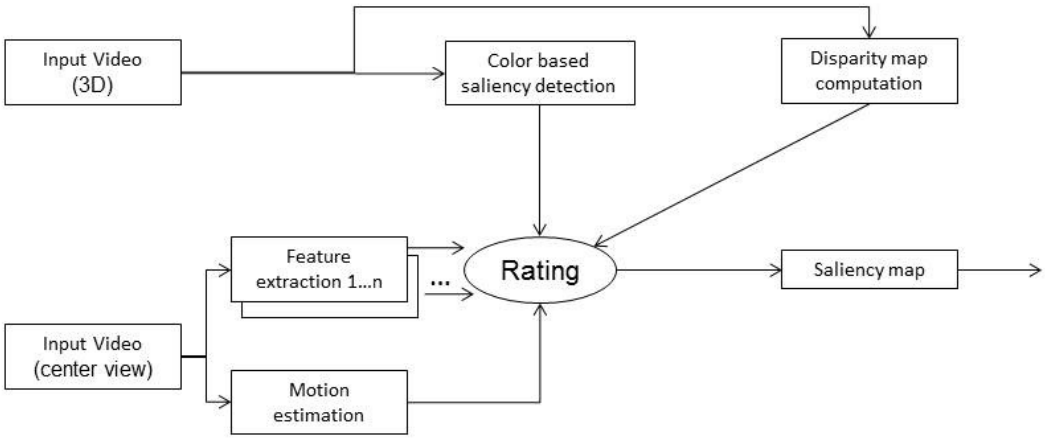


Figure 31: First diagram version of proposed system.

7.3 Content Format Adaptation

Content format adaptation is in some aspects a quite critical operation because the original image formation of the cinematographer is modified. For this reason it is not only a demanding technical problem, it also might result in right conflicts. Nevertheless content format adaptation can be a very useful application that improves the users viewing experience. Content format adaptation becomes necessary when video and display do not fit together – either in resolution or in aspect ratio. In the ROMEO project we will only focus on aspect ratio adaptation. Resolution issues might be necessary in some cases as well but we believe the rapidly increasing panel resolution even in small devices will make this topic less important in the future.

Aspect ratio adaptation on the other hand we believe will always be required. On the production side as well as on the display side there are many different aspect ratios. They reach from the old 4:3 format (still present in Broadcaster's archives), to 3:2 (iPhone), 16:9 (standard TV production) and 21:9 (some cinema productions), just to name a few. The simplest method to adapt a video to a non-fitting display is to add letter or pillar boxes. On large screen this might be an acceptable solution, on small portable devices this method often leads to tiny images being displayed. In this case a crop and rescale of the video might improve the viewing experience. Previous investigations show that users prefer cropping solutions [51] at least in some cases. When a video is cropped visual areas are cut off indeed, but the complete screen is filled with active video. One has to make sure of course that no important image content is cut off.

In ROMEO we intend to build a demonstrator that meets this requirement. The cropping concept is based on the salient map that is described in the previous chapter. But in addition further aspects needs to be considered when it comes to 3D cropping. Furthermore we introduce some user defined parameters in order to allow some basic adjustments of the algorithm. Figure 32 gives an overview of the proposed concept.

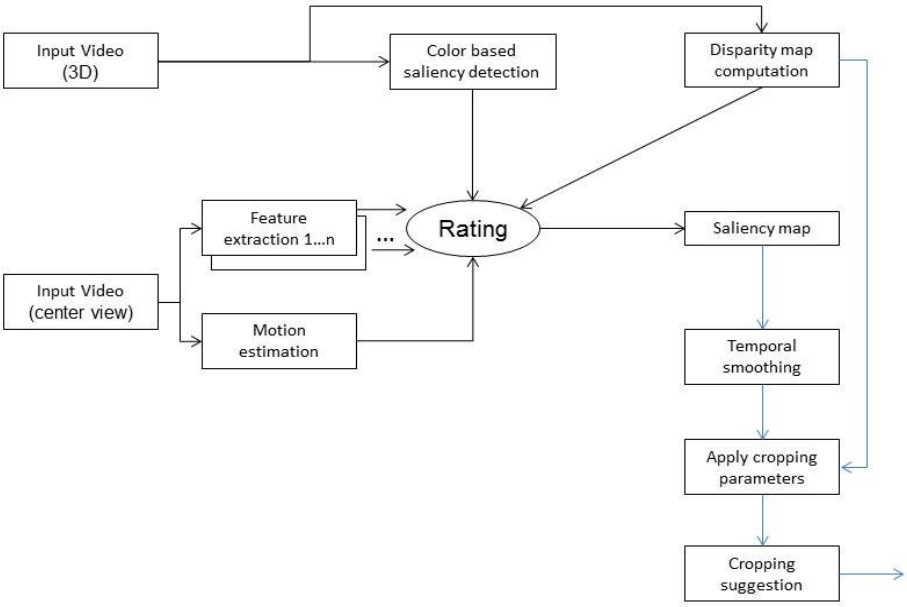


Figure 32: Suggested cropping concept.

With the resulting saliency map we calculate first cropping suggestion. Obviously it is desirable that no interesting object from the feature map is cut off. However the exact algorithm still needs to be defined and will be described in D3.5. This first suggestion is then checked against the parameters described below.

7.3.1 Smooth Motion

The first cropping suggestion is refined in order to maintain smooth camera motion. Since dynamic cropping can influence camera motion we want to ensure that motion is still smooth after the cropping process. Here it must be considered that camera motion and cropping motion add up to the overall motion. To preserve smoothness in camera movement a low pas filter is applied to the first cropping suggestion.

A second aspect in this context is local extrema preserving. Assuming that we have to crop left and /or right, i.e. horizontal panning is affected. Let $f_{cam}(t)$ be the function that describes the horizontal camera movement starting with $f_{cam}(t_0) = 0$. And let $f_{crop}(t)$ be the function that describes deflection of the cropped area from the center.Then the overall horizontal movement $f(t)$ can be written as

$$f(t) = f_{cam}(t) + f_{crop}(t)$$

With local extrema preserving we want $f(t)$ to have the same extrema as $f_{cam}(t)$. Local extrema of $f_{cam}(t)$ are those points where panning direction is changed. Those extrema must not be affected by cropping. Additionally further extrema must not be added. An additional local extrema would mean to change panning direction at a point where original camera movement keeps panning.

To estimate the camera movement in order to retrieve $f_{cam}(t)$ we intend to use the methods described in [51].

7.3.2 3D constraint

With 3D video content another aspect needs to be considered. It will also be handled in the second step.

When objects that do not lie within the display layer are cut off at the left or right edge, the viewer's eyes get inconsistent information. In this case one eye sees more of the object than the other one. When the object lies behind the display it is like looking out of a window. The human's visual system is used to that. But if it is in front of the display it usually irritated the viewer and can cause fatigue or distort the 3D effect [52]. This artifact is known as "window violation issue". In a quality production it is a well-known rule that objects in front of a screen are cut off neither left nor right. However when cropping is applied this rule must be considered as well. Therefore the disparity map is analyzed to protect all front objects from being cut off. A detailed description will follow in Deliverable D3.5.

7.3.3 Cropping parameter definition

Before the video is finally cropped some parameters that can be set by the user will be considered. They partly overlap with cropping constraints that are described above but this time they can be controlled manually. This gives an operator the possibility to adjust the cropping application to its need. Two parameters were defined.

Maximum window velocity defines how fast the cropping window will move at most. Value 0 means that the cropping window must be at one position throughout the complete scene. Value -1 defines that no limit is set.

Maximum deflection defines how far the window can deflect from the center position. The value is set in image width.

Suggestions for reasonable values will be defined empirically. These parameters should be set by the broadcaster or content producer when cropping is applied. It is most likely that different platforms vary in their requirements and concepts when it comes to cropping.

7.4 Evaluation and demonstrator

For evaluation of the saliency map quality an eye tracking system is engaged. This system allows tracking of viewer's eyes and provides this data for analysis. We will run tests were viewers are asked to watch a set of test videos. The eye tracker provides accurate data of the viewer's gaze point. This data allows estimating which regions in a video are salient for the viewer. The results will be compared to our saliency maps.

To demonstrate our automatic aspect ratio adaptation two scenarios have been chosen:

- Adapt 16:9 content to 3:2 screens. In this scenario the video is cut off left and/or right to fit the 3:2 aspect ratio. We believe it is a relevant example since 16:9 is used for HDTV productions and 3:2 (being the aspect ratio of the iPhone) is a widely spread screen size.
- Adapt 4:3 content to 16:9. This scenario is less relevant for ROMEO application but quite important for broadcasters. When broadcasting old 4:3 content from archive pillar boxes are generally inserted. But for online distribution a crop and rescale automatism could be an interesting alternative. Precisely because it also addresses portable devices.

The automatic adaptation will be applied to different example videos. The adaptation described above will be evaluated and compared to manual cropping as well as to letter box and pillar box.

Details for both evaluation and demonstrator will be described in D3.5.

8 CONCLUSIONS AND FUTURE WORK

This deliverable describes the initial investigations of the ROMEO consortium on QoE and VA Modelling work. This describes in details the audio-visual subjective experiments performed and the analysis of those subjective results.

In term of QoE modelling for compression artefacts in stereoscopic video, psychophysical experiments were performed to identify and isolate artefacts and their thresholds that affect the user perception. A subjective experiment was performed to analyze the effect of compression artefacts produced by H.264/AVC and HEVC video codecs. The existing metrics were compared for correlation with the subjective scores and an initial metric based on weighted VQM was presented in this deliverable. This deliverable also describes the initial quality metric developed to assess the quality of depth maps in term of its rendering capability. Further subjective assessments with more videos and improvement of the quality metrics will be performed and reported in the future.

A preliminary QoE model for network related impairments (packet loss) as well as physical layer losses, which is based on objective 2D metrics, is proposed. A set of experiments with stereoscopic video sequences indicate a close correlation between the proposed QoE model and the measured 3D MOS. The aim of future research is to explicitly model the impact of specific networking parameters (packet loss, delay, Unequal Error Protection, Forward Error Protection, Mobility) on the perceived quality of both stereoscopic and color-depth sequences. An optimization (maximization or minimization) scheme will be proposed for optimizing perceived 3D video quality during streaming over erroneous transmission channels (i.e considering parameters of both physical and network/transport layers).

The relationship between the perceived spatial audio QoE and the physical parameters of multichannel audio has been introduced. A review of the state-of-the-art in QoE measurements of Audio Systems have been presented in this deliverable. The introduced prediction model will be improved by modifying different parameters of the model and validated through another set of subjective experiments, including the ROMEO media contents, and a conclusive spatial audio QoE prediction model will be suggested in the next deliverable on QoE.

In the context of VAM, it is proposed to detect the most plausible salient regions by combining the two state-of-the-art visual saliency approaches. For further video encoding, depth based saliency map is presented to detect important regions of interest. A rating mechanism is explained which will decide how to weight those maps, that leads to a final salient map. Finally, a cropping concept is also shown which will use the resulting saliency map in conjunction with some user define parameters for cropping suggestion.

In the final deliverable, it is planned to demonstrate a framework which will consider video as an input. Proposed saliency detection methods (with and without using depth information) will be applied on the video to detect salient regions. Then using the results of salient regions cropping will be performed and the results will be shown on mobile and display devices in order to demonstrate the automatic aspect ratio adaptation. In addition to that, an evaluation of the saliency map quality with the eye tracking system result will also be shown at the end.

9 REFERENCES

[1] B. Julesz, 'Cyclopean perception and neurophysiology', *Investigative Ophthalmology & Visual Science*, vol. 11, no. 6, p. 540, 1972.
[2] M. G. Perkins, 'Data compression of stereopairs', *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684–696, Apr. 1992.
[3] W. D. Reynolds and R. V. Kenyon, 'The wavelet transform and the suppression theory of binocular vision for stereo image compression', in , *International Conference on Image Processing, 1996. Proceedings*, 1996, vol. 1, pp. 557–560 vol.2.
[4] L. Lei and C. M. Schor, 'The spatial properties of binocular suppression zone', *Vision research*, vol. 34, no. 7, pp. 937–947, 1994.
[5] H. Kalva, L. Christodoulou, L. M. Mayron, O. Marques, and B. Furht, 'Design and evaluation of a 3D video system based on H.264 view coding', in *Proceedings of the 2006 international workshop on Network and operating systems support for digital audio and video*, Newport, Rhode Island, 2006, pp. 12:1–12:6.
[6] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, 'Subjective study on compressed asymmetric stereoscopic video', in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 4021–4024.
[7] G. Saygili, C. G. Gurler, and A. M. Tekalp, 'Evaluation of Asymmetric Stereo Video Coding and Rate Scaling for Adaptive 3D Video Streaming', *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 593–601, Jun. 2011.
[8] G. Saygili, C. G. Gürler, and A. M. Tekalp, 'Quality assessment of asymmetric stereo video coding', in *2010 17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 4009–4012.
[9] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, 'New full-reference quality metrics based on HVS', in *CD-ROM Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, 2006.
[10] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, 'On between-coefficient contrast masking of DCT basis functions', in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2007.
[11] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, '3D-DCT based perceptual quality assessment of stereo video', in *2011 18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2521 –2524.
[12] A. Benoit, P. L. Callet, P. Campisi, and R. Cousseau, 'Quality Assessment of Stereoscopic Images', *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 659024, Jan. 2009.
[13] J. Seo, X. Liu, D. Kim, and K. Sohn, 'An Objective Video Quality Metric for Compressed Stereoscopic Video', *Circuits, Systems, and Signal Processing*, vol. 31, no. 3, pp. 1089–1107, Nov. 2011.
[14] A. M. Treisman and G. Gelade, 'A Feature-Integration Theory of Attention.', *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.
[15] A. Treisman, 'Features and objects in visual processing', *Sci. Am.*, vol. 255, no. 5, pp. 114–125, Nov. 1986.
[16] X. Hou and L. Zhang, 'Saliency detection: A spectral residual approach', in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, 2007, pp. 1–8.
[17] X. Cui, Q. Liu, and D. Metaxas, 'Temporal spectral residual: fast motion

saliency detection', in *Proceedings of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, pp. 617–620.
[18] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, 'Salient region detection and segmentation', in *Proceedings of the 6th international conference on Computer vision systems*, Berlin, Heidelberg, 2008, pp. 66–75.
[19] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk, 'Frequency-tuned salient region detection.', in *CVPR*, 2009, pp. 1597–1604.
[20] R. Achanta and S. Süsstrunk, 'Saliency detection using maximum symmetric surround.', in *ICIP*, 2010, pp. 2653–2656.
[21] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, 'Global contrast based salient region detection', in *CVPR*, 2011, pp. 409–416.
[22] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, 'Unsupervised extraction of visual attention objects in color images', *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 16, no. 1, pp. 141–145, Sep. 2006.
[23] B. C. Ko and J.-Y. Nam, 'Object-of-interest image segmentation based on human attention and semantic region clustering', *J. Opt. Soc. Am. A*, vol. 23, no. 10, pp. 2462–2470, Oct. 2006.
[24] N. Jacobson, Y.-L. Lee, V. Mahadevan, N. Vasconcelos, and T. Q. Nguyen, 'A novel approach to FRUC using discriminant saliency and frame segmentation', *Trans. Img. Proc.*, vol. 19, no. 11, pp. 2924–2934, Nov. 2010.
[25] N. Jacobson and T. Q. Nguyen, 'Video processing with scale-aware saliency: Application to Frame Rate Up-Conversion.', in *ICASSP*, 2011, pp. 1313–1316.
[26] C. Vu and D. M. Chandler, 'Main subject detection via adaptive feature refinement', *Journal of Electronic Imaging*, vol. 20, no. 1, Mar. 2011.
[27] C. Koch and S. Ullman, 'Shifts in selective visual attention: towards the underlying neural circuitry.', *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.
[28] X. Hou, J. Harel, and C. Koch, 'Image Signature: Highlighting Sparse Salient Regions.', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.
[29] M. J. Collins and A. Goode, 'Interocular blur suppression and monovision', *Acta Ophthalmologica*, vol. 72, no. 3, pp. 376–380, Apr. 1995.
[30] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, A. Fernando, S. Dogan, A. Kondoz, and S. Savas, 'Psycho-physical limits of interocular blur suppression and its application to asymmetric stereoscopic video delivery', in *Packet Video Workshop (PV), 2012 19th International*, 2012, pp. 184 –189.
[31] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, 'A no-reference perceptual blur metric', in *2002 International Conference on Image Processing. 2002. Proceedings*, 2002, vol. 3, p. III–57 – III–60 vol.3.
[32] H. long Kim and S. G. Choi, 'A study on a QoS/QoE correlation model for QoE evaluation on IPTV service', in *Proceedings of the 12th international conference on Advanced communication technology*, Piscataway, NJ, USA, 2010, pp. 1377–1382.
[33] C. Tselios, M. Tsagkaropoulos, I. Politis, and T. Dagiuklas, 'Valuing quality of experience: A brave new era of user satisfaction and revenue possibilities', in *FITCE Congress (FITCE), 2011 50th*, 2011, pp. 1 –6.
[34] D. Soldani, 'Means and methods for collecting and analyzing QoE measurements in wireless networks', in *World of Wireless, Mobile and Multimedia Networks, 2006. WoWMoM 2006. International Symposium on a*, 2006, p. 5 pp. –535.
[35] A. Talari and N. Rahnavard, 'Unequal error protection rateless coding for efficient MPEG video transmission', in *Military Communications Conference, 2009.*

MILCOM 2009. IEEE, 2009, pp. 1 –7.
[36] L. Rizzo, 'Dummynet: a simple approach to the evaluation of network protocols', *SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 1, pp. 31–41, Jan. 1997.
[37] 'ITU-R Recommendation BT.500-11 Methodology for the subjective assessment of the quality of television pictures'. 2002.
[38] MPEG WG11 SC29, *View Synthesis Based on Disparity/Depth Software*. 2008.
[39] MUSCADE Consortium, *Specification of A/V rendering and display adaptation- Phase II*. MUSCADE Project Deliverable 1.4.3, 2011.
[40] MUSCADE Consortium, *3D capture and post production design –Phase I*. MUSCADE Project Deliverable 2.2.1, 2010.
[41] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, 'Towards a New Quality Metric for 3-D Synthesized View Assessment', *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332 –1343, Nov. 2011.
[42] S. George, 'Objective models for predicting selected multichannel audio quality attributes', University of Surrey, Guildford, 2009.
[43] Thiede, T.T., William C.; Bitto, Roland; Schmidmer, Christian; Sporer, Thomas; Beerends, John G.; Colomes, Catherine, 'PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality', *Journal of Audio Engineering Society*, vol. 48, no. 1, pp. 3–29, 2000.
[44] Zielinski, Slawomir K.; Rumsey, Francis; Kassier, Rafael; Bech, Søren, 'Comparison of Basic Audio Quality and Timbral and Spatial Fidelity Changes Caused by Limitation of Bandwidth and by Down-mix Algorithms in 5.1 Surround Audio Systems', *Journal of Audio Engineering Society*, vol. 53, no. 3, pp. 174–192, 2005.
[45] Bates, E., 'Monophonic source localization for a distributed audience in a small concert hall.', *10th Int. Conference on Digital Audio Effects (DAFx-07). 2007. Bordeaux, France*.
[46] A. Field, *Discovering Statistics Using SPSS. 3rd ed. Introducing Statistical Methods.* SAGE Publications Ltd, 2009.
[47] J. Roden, J. Breebart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Holzer, *A Study of the MPEG Surround Quality Versus Bit-rate Curve*. New York, USA: , 2007.
[48] J. Herre and others, 'MPEG Surround - The ISO/MPEG standard for efficient and compatible multichannel audio coding', *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.
[49] D. Marston, F. Kozamernik, G. Stoll, and G. Spikofski, *Further EBU Test of Multichannel Audio Codecs*. Munich, Germany: , 2009.
[50] ITU-R, *Method for Objective Measurements of Perceived Audio Quality*. 2001.
[51] J. Deigmoeller, 'Intelligent image cropping', 2011.
[52] W.-N. M. Peter Kauff, Ralf Schäfer, Frederik Zilly, Josef Kluger, 'Stereoscopic Analyzer (STAN)', 2010.

APPENDIX A: GLOSSARY OF ABBREVIATIONS

A	
AEW	Average Edge Width
AVC	Audio Video Correlation
B	
BAQ	Basic Audio Quality
C	
CSVQ	Compressed Stereoscopic Video Quality Metric
D	
DCT	Discrete Cosine Transform
DIBR	Depth Image Based Rendering
DMOS	Differential Mean Opinion Scores
DMQ	Depth Map Quality Metric
F	
FIT	Feature Integration Theory
FSF	Front Spatial Fidelity
G	
GS	Guided Search
H	
HEVC	High Efficiency Video Coding
HP	High Priority
HRTF	Head-Related Transfer Functions
HVS	Human Visual System
I	
IBS	Interocular Blur Suppression
IS	Signature Approach
J	
L	
LP	Low Priority
M	
MSE	Mean Squared Error
O	
ODG	Objective Difference Grade
P	
P2P	Peer-to-Peer
PEAQ	Perceived Audio Quality
PSNR	Peak Signal to Noise Ratio
PSPNR	Peak Signal to Perceptual Noise Ratio
Q	
QoE	Quality of Experience

QoS	Quality of Service
QP	Quantization Parameter
S	
SR	Spectral Residual
SSF	Spatial Fidelity
T	
TIS	Temporal Image Signature
TSR	Temporal Spectral Residual
U	
UEP	Unequal Error Protection
V	
VAM	Visual Attention Modelling
ViSBD	View Synthesis Reference Software
W	
WP	Work Package