Modelling of the Stereoscopic HVS

Atanas Boev  Maija Poikela  Atanas Gotchev  Anil Aksay
Modelling of the Stereoscopic HVS

Atanas Boev, Maija Poikela, Atanas Gotchev, Anil Aksay

Abstract: The report presents an overview of fields of study, related with modelling of the human visual system and perception of visual information – HVS anatomy and physiology, HVS modelling, and visual quality metrics. State-of-the-art in HVS modelling, including modelling of light adaptation, contrast sensitivity, and spatiotemporal masking is presented. Special attention to the visual mechanisms responsible for stereoscopic perception is paid. The algorithms behind the most popular visual quality metrics are described along with a comparative study of 15 metrics over a database of stereoscopic video streams. Arguments are provided about why 2D metrics fail when applied to 3D quality evaluation. The overview grounds the need of running subjective tests about the relative importance of different 3D visual cues on portable autostereoscopic display and about the influence of varying depth on the overall visual quality. The results of first stage of these studies are presented. Finally, a model of the stereoscopic HVS, suitable for feature based 3D video quality evaluation and the corresponding image processing channel are discussed.

Keywords: 3DTV, stereoscopic quality, mobile video, HVS modelling, quality estimation, 3D quality metrics
Executive Summary

This work aims at creating a model of the stereoscopic human visual system (HVS), which to be used later on for perceptual quality estimation of 3D video streams. Such model is an outcome of a multidisciplinary study, which combines knowledge about HVS anatomy and physiology, psychophysical experiments, models of HVS features, an overview of existing visual quality metrics as well as results of specifically designed and conducted subjective tests.

An overview of the anatomy of the eye, optical nerve and areas of the brain cortex, which are dealing with vision, is given first. The continuous process of visual perception and the specific properties of vision – light adaptation, contrast sensitivity, colour and motion perception are described. Special attention is paid to the visual mechanisms which facilitate stereoscopic perception.

An overview of the state-of-the-art in HVS modelling is presented. Models of various HVS properties are discussed – such as contrast sensitivity function, perceptual colour spaces, temporal and spatial masking models. Additionally, different approaches in HVS modelling are discussed, these including single-channel and multi-channel models, and psychophysical and neurobiological modelling.

Furthermore, a state-of-the-art review of visual quality metrics is provided and the algorithms behind the most popular ones are described. A comparative study of 15 metrics and their performance on a database of impaired 3D videos has been accomplished. Eventually, arguments about why 2D metrics are not sufficient to register (predict) the 3D video quality are given. These are taken into account for the design and execution of subjective tests to quantify the relative importance of 3D visual cues on an auto-stereoscopic display.

Finally, a HVS model, which mimics the stages of stereoscopic visual perception and an image processing channel for stereoscopic quality estimation are presented. The latter is to be used in the design of a feature-based objective metric based also extensively on the results of the recently-accomplished and ongoing subjective tests.
# Table of Contents

1. Introduction .................................................................................................................. 5
   1.1 Motivation ............................................................................................................. 5
   1.2 Outline .................................................................................................................. 5

2. Physiology of vision ...................................................................................................... 6
   2.1 Anatomy of the visual system ............................................................................. 6
      2.1.1 Eye ............................................................................................................... 6
      2.1.2 Visual pathways ............................................................................................ 7
      2.1.3 Visual cortex ............................................................................................... 7
   2.2 Visual perception .................................................................................................... 8
      2.2.1 Light adaptation ........................................................................................... 8
      2.2.2 Contrast sensitivity ..................................................................................... 8
      2.2.3 Spatial vision ............................................................................................... 9
      2.2.4 Perception of colour ................................................................................... 9
      2.2.5 Perception of motion .................................................................................. 10
   2.3 Eye movements ...................................................................................................... 10
      2.3.1 Versions and vergences ............................................................................. 10
      2.3.2 Saccadic movements .................................................................................. 11
   2.4 Stereopsis .............................................................................................................. 11
      2.4.1 Binocular geometry ..................................................................................... 11
      2.4.2 Binocular stereopsis ................................................................................... 12
      2.4.3 Binocular depth cues ................................................................................. 12
      2.4.4 Convergence-accommodation rivalry ......................................................... 13

3. Modelling of human visual system .......................................................................... 14
   3.1 HVS modelling overview .................................................................................... 14
   3.2 Single and multichannel models ......................................................................... 15
   3.3 Models of HVS properties ................................................................................ 15
      3.3.1 Luminance .................................................................................................. 15
      3.3.2 Colour ........................................................................................................ 16
      3.3.3 Contrast sensitivity function ..................................................................... 17
      3.3.4 Temporal and spatial masking .................................................................. 18
   3.4 Modelling binocular vision ................................................................................ 19

4. Visual quality metrics ............................................................................................... 21
   4.1 Model-based and feature-based approach ......................................................... 21
   4.2 Metric classification ............................................................................................ 22
   4.3 Overview of visual quality metrics ................................................................... 23
      4.3.1 PSNR ......................................................................................................... 23
      4.3.2 SSIM .......................................................................................................... 23
      4.3.3 VSSIM ....................................................................................................... 24
      4.3.4 CW-SSIM ................................................................................................. 24
      4.3.5 ICIQ ........................................................................................................ 24
      4.3.6 VSNR ....................................................................................................... 25
      4.3.7 VQM ........................................................................................................ 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.8</td>
<td>Sarnoff’s JND</td>
<td>26</td>
</tr>
<tr>
<td>4.3.9</td>
<td>DVQ</td>
<td>28</td>
</tr>
<tr>
<td>4.3.10</td>
<td>PDM</td>
<td>29</td>
</tr>
<tr>
<td>4.3.11</td>
<td>PQS</td>
<td>30</td>
</tr>
<tr>
<td>4.3.12</td>
<td>PSNR-HVS-M</td>
<td>30</td>
</tr>
<tr>
<td>4.3.13</td>
<td>VIF</td>
<td>31</td>
</tr>
<tr>
<td>4.4</td>
<td>Quality metric evaluation</td>
<td>31</td>
</tr>
<tr>
<td>4.4.1</td>
<td>VQEG studies</td>
<td>31</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Performance parameters</td>
<td>32</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Tests sets</td>
<td>32</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Comparison of 2D quality metrics for 3D video artefacts</td>
<td>33</td>
</tr>
<tr>
<td>4.5</td>
<td>Stereoscopic quality metrics</td>
<td>38</td>
</tr>
<tr>
<td>4.6</td>
<td>Why 2D metrics fail for 3D images</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>Relative importance of depth cues</td>
<td>42</td>
</tr>
<tr>
<td>5.1</td>
<td>Research method</td>
<td>42</td>
</tr>
<tr>
<td>5.2</td>
<td>Test content</td>
<td>42</td>
</tr>
<tr>
<td>5.3</td>
<td>Subjective tests</td>
<td>45</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Test design</td>
<td>45</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Test procedure</td>
<td>45</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Simulator Sickness Questionnaire</td>
<td>47</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Test automation</td>
<td>48</td>
</tr>
<tr>
<td>5.4</td>
<td>Results</td>
<td>48</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Speed of Depth Estimation</td>
<td>48</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Correctness of depth estimation</td>
<td>50</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Quality Ratings</td>
<td>51</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Acceptance Ratings</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>Towards stereoscopic HVS modelling for feature-based 3D video quality evaluation</td>
<td>54</td>
</tr>
<tr>
<td>6.1</td>
<td>Relative impact of 2D and 3D depth cues on overall 3D perception</td>
<td>54</td>
</tr>
<tr>
<td>6.2</td>
<td>Impact of binocular disparity on overall quality</td>
<td>54</td>
</tr>
<tr>
<td>6.3</td>
<td>Image processing channel for stereoscopic quality estimation</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>Conclusions</td>
<td>59</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Motivation

One of the goals of MOBILE3DTV project is to develop algorithms for objective assessment of the subjective quality of stereoscopic video content. Quality assessment algorithms analyse the video content, and try to predict its perceptual quality. In order to design such algorithms, one needs to study the characteristics and limits of the human visual system (HVS), and to understand the process by which HVS combines the visual information of the two eyes and perceives a scene in depth.

The knowledge of how HVS operates comes from a variety of scientific fields – anatomy, physiology, psychology. While a wide range of HVS features, including binocular perception, are thoroughly studied, models of stereoscopic perception are still scarce, and in most cases – simplistic. Most common approach is to estimate the quality seen by each eye separately, and to combine the two measurements into one compound quality metric. This approach would fail to predict the effects of binocular masking and facilitation on the overall perceived quality.

In this report, we aim to create a model of the stereoscopic HVS, which to be used for perceptual quality estimation of 3D video streams. We combine knowledge of four fields of study – 1) HVS anatomy and physiology, which study how HVS operates; 2) psychophysical studies and HVS modelling, which study how to represent the visual process by a mathematical model; 3) visual quality metrics, which answer the question how to use a combination of HVS models for reliable estimation of the perceptual quality, and 4) subjective tests which help in perceptually quantifying the effect of different 3D visual cues and tuning the parameters of the metric under development.

1.2 Outline

Chapter 2 begins with an overview of the anatomy of the eye, optical nerve and areas of the brain cortex, which are dealing with vision. Then, the continuous process of visual perception is examined – light adaptation, contrast sensitivity, perception of colour and motion and eye movements. Chapter 2 ends with discussion of the visual mechanisms which allow stereoscopic perception.

Chapter 3 deals with the state-of-the-art in HVS modelling including contrast sensitivity function, and colour models.

Chapter 4 is about visual quality metrics. A classification of metrics is given, and the most popular metrics of visual quality are described. We study how the performance of quality metrics is evaluated, and the available databases of test images and video streams. We present the VQEG studies on quality metric performance and a performance evaluation of the non-proprietary algorithms for visual quality estimation. Chapter 4 ends with a discussion about why 2D quality metrics are not sufficient to evaluate the quality of 3D video.

Chapter 5 describes subjective quality experiments, which aim at estimating the relative importance of 2D and 3D depth cues over the overall perceived quality, and the ability of the observer to make fast and accurate depth assessments.

Finally, Chapter 6 explains our approach for objective estimation of stereoscopic quality. Based on the findings of the above subjective tests, a new set of tests has been designed and is currently being conducted. The ultimate goal is to combine the knowledge on stereopsis with models of the 2D vision and the outcomes from the subjective tests. As a result, we propose an image processing channel for stereoscopic quality estimation.
2 Physiology of vision

2.1 Anatomy of the visual system

2.1.1 Eye

The optical system of the human eye is composed of the cornea, the aqueous humor, the lens, and the vitreous humor [3], [5]. The lens and cornea together provide two thirds of the total optical power of the eye [2]. The cornea has a greater optical power than the lens but that of the lens can be modified to focus at different distances.

Light enters the eye through a hole in the middle of the iris, called pupil. The size of the pupil is regulated by the iris. The lens has a convex shape and the light refracts as it enters it. Muscles attached to the lens can change the curvature of it to accommodate to objects at different distances. The images are brought into focus on the retina, which is a neural tissue at the back of the eye, consisting of three layers of cell bodies and two layers containing the synaptic interconnections between the neurons. Because of the optical characteristics of the eye, the images projected onto the retina are reversed. The image is thoroughly pre-processed in the retina before it is passed to other parts of the brain.

Figure 2.1: Human eye, source: Wikimedia Commons [29]

Light entering the retina on most regions traverses several layers of neurons, including ganglion cells, amacrine cells, bipolar cells and horizontal cells. Ganglion cells can be classified in two different groups: parvocellular cells, which response to fine image details and chromatic information encoding, and magnocellular cells, which response to form, motion, depth and small differences in light level. These cells process the photoreceptor signals. After this the light goes
through the light-sensitive photoreceptors (cones and rods) and finally is absorbed in the pigment layer. An interconnection between these cells is called a receptive field. The cones are specialized to detecting colours and function in bright light. The rods don’t take part in colour vision and function mainly in low illumination levels. The interconnections and functions of these cells are more broadly introduced in [1].

The area of sharp focus in the retina is called fovea, and in this area the cell bodies are shifted to the side enabling higher resolution by a smaller distortion caused by passing several layers of cells. The central fovea contains the most of cones but no rods at all, and while going further from the central fovea the proportion and density of cones decrease and that of the rods increase. The cone-rich region has a high acuity vision in bright light, whilst the rod-rich area has a good sensitivity in dim light. The size and spacing of the photoreceptors specify the maximum spatial resolution of the human visual system. There are approximately 127 million receptors in the retina (120 million rods and 7 million cones), but only one million ganglion cells in the optic nerve. A signal from a single cone in the fovea is sent to a bipolar cell and further to ganglion cells, whereas in the periphery many rods may connect to a single bipolar cell. For more detailed information about the anatomy of the human eye, see [1], [3], [4] and [21].

2.1.2 Visual pathways

The visual pathways carry visual information from the ganglion cells of the retina to the brain. There is a blind spot in the eye in the location where the optic nerve leaves the retina, but it can not be easily detected, since it is not located in the same spot at the both eye’s visual field.

The nasal side of the visual field is projected on the temporal side of the retina and the temporal side of the visual field on the nasal side of the retina. In the optical nerve the optical tracts coming from the nasal sides of the eyes cross at the optic chiasm, and the optical tracts from the temporal sides lead straight to corresponding temporal lobes [18]. In the optical chiasm the crossing fibres are rearranged. The fibres then join to form optic tracts and pass by the way of the optic radiation to the visual cortex in the brain. The right side of the brain receives information from only the left side of the visual field and vice versa. This allows the images of an object on the right and left retinae to be processed in the same part of the brain. The optic nerve, as well as retina, is considered a part of the central nervous system [5].

There is a lateral geniculate nucleus (LGN) located on both the hemispheres of the brain, each of which receives the input from the retina from one half of the visual field. There are approximately one million neurons in six layers comprising the LGN. Different layers receive input from different types of cells.

Moreover, LGN controls how much of the information is allowed to pass. This is done with the help of feedback from the primary visual cortex. The feedback suggests that the LGN has an even more important function than what has been thought until now [21].

2.1.3 Visual cortex

Visual cortex is located at the back of the brain. The signals come from lateral geniculate nucleus and arrive at primary visual cortex (V1), which is the largest part of the human visual system. There is an enormous variety of cells in the visual cortex. These cells respond to different kind of stimuli, like particular frequencies, colours or direction.

The cortical area V1 consists, like the cortex in general, of different layers. The superficial layer 1 has only few neurons but many axons, dendrites, and synapses. These are collectively called neuropil. Layers 2 and 3 have a dense array of cell bodies and many local dendritic interconnections. The layers 1-3 are called the superficial layers of the cortex and they receive the input from koniocellular layers. Layer 4 has been further divided to several parts, labelled 4A,
4B, 4Cα, and 4Cβ. Layer 4C receives the primary input from the parvocellular and magnocellular layers of the LGN [4]. Layer 5 sends a major output to the superior colliculus, which is a structure in the midbrain. Layer 6 is dense with cells and sends output back to the LGN. The signals coming to the V1 are complex and very specific, which suggests that the interconnections within the area V1 are also highly specific [1].

From the primary visual cortex the information is sent further to other locations of the brain through two primary pathways, called dorsal stream and ventral stream. The former begins with V1 and goes to the area V2 and dorsomedial area (also known as V6) and visual area MT (V5). The latter, the ventral stream, goes from the area V1 through V2 and area V4 and finally to inferior visual cortex [19]. The dorsal stream is associated with motion, representation of object locations, and control of the eyes and arms, especially when visual information is used to guide saccades or reaching. The ventral stream is associated with form recognition and object representation. It is also associated with storage of long-term memory [3], [21].

2.2 Visual perception

There are considerable variations in the optics between individuals and an eye also goes through continuous changes throughout the life. Thus, when modelling the human eye approximations have to be made.

2.2.1 Light adaptation

The human visual system is capable of adapting to a great range of light intensities. There are three mechanisms for light adaptation, which are presented below.

The pupillary aperture size regulated by the iris influences the amount of light entering the eye. The adaptation happens in a matter of seconds. The muscles attached to the iris react to differences in illumination level but also to sympathetic impulses, such as shock. In high illumination levels the pupil diminishes, and in low illumination levels widens. In darkness the enlargement of the pupil can increase the amount of light entering the eye by as much as 15 to 30 times.

The photoreceptors (rods and cones) have a chemical mechanism to adapt to light. The rods are responsible for scotopic and cones for photopic vision, latter reacting to high light levels. When the light intensity rises, the concentration of photochemicals in the receptors decreases. This makes them less sensitive. In low light levels the concentration of rod photochemicals, called rhodopsin, rises together with the sensitivity of the receptors. This process can take up to several hours. The more the receptors have visual pigment, the more likely it is for a photon to hit the pigment and evoke a photochemical reaction.

An adaptation mechanism faster than the one described above is adaptation at the neural level. This mechanism works by increasing or decreasing the signal output of retinal neurons. Some ganglion cells gather impulses from a larger amount of rod cells, thus making the sensitivity to detail weaker but light sensitivity stronger.

Adaptation to bright light after dim lighting happens much faster than vice versa. When entering a higher illumination level, the pupils become smaller restricting the amount of light entering the eye. The light falling on the retina breaks down most of the visual pigments. This irritates the light receptors and causes dazzling [21].

2.2.2 Contrast sensitivity

Rods and cones convert light energy into signals that can be interpreted by the brain. There is a challenge at the encoding of the visual signal, because the ambient light intensity, from a dim
evening to bright sunny day, can vary over six orders of magnitude. An individual neuron has a much smaller response range – only of two to three orders of magnitude. The solution to this is encoding of the local contrast rather than absolute image level [3]. The changing of neural and behavioural responses as a function of mean background intensity is called visual adaptation.

The human visual system is much more sensitive to relative differences in luminance than the absolute luminance level. The image contrast is the ratio of the local intensity and the average image intensity [3]. The minimum contrast necessary for an observer to detect a change in intensity is called a threshold contrast. In order to represent the image contrast the neurons in the visual pathway compensate for changes in the mean illumination level [4], [21].

2.2.3 Spatial vision

The size and spacing of the retinal photoreceptors (rods and cones) determine the maximum spatial resolution of the human visual system [4]. In the central retina, fovea, there is a huge concentration of cones, whereas there are hardly any rods at all. As the distance from the fovea increases, the amount of rods increases and that of cones decreases. In these peripheral areas the cones are of a bigger size than in the fovea. The location of rods also explains why it's sometimes easier to see objects in dim lighting by looking slightly past the object. In the fovea a single receptor can have its own bipolar and ganglion cells, which makes the ability to detect fine details good. Half of the brain’s visual area processes the information coming from the fovea.

In other locations of the retina signals from many receptors converge onto a single neuron. This reduces the resolution and makes visual acuity poor. On the other hand the light sensitivity is good in these parts since the rods are very sensitive light detectors and sample the retina very finely.

Masking occurs when a stimulus visible by itself cannot be seen because of the presence of another stimulus. Masking is strongest when the stimuli have similar characteristics, i.e. colour, frequency or orientation [21]. The results from masking and adaptation experiments were the major motivation for developing a multi-channel theory of vision [4].

According to [3], there is a collection of neurons whose responses capture the image information that is available to the observer. Collections of neuron responses together make up neural images, differing from each other by the location of the centre of receptive field. These neural images can be used to visualize the neural response of an input image. The pattern sensitivity can be predicted using this kind of neural images.

2.2.4 Perception of colour

The appearance of an object’s colour depends most on its surface-reflectance properties [3]. The computational analyses of colour help in understanding the colour perception: the neural responses to wavelength. The appearance of a particular colour depends on the viewer’s state of adaptation, both globally and locally, the size, configuration, and location of the stimulus in the visual field, the colour and location of other objects in the scene, the color of the background and surround, and the colours of objects presented after the one in question.

The human eye can detect light at wavelengths 397-723 nm. Early colour perception happens in the retina, where the light-sensitive cones react to different wavelengths of light. These cones are called L-cones, M-cones and S-cones originating from their ability to react to long, medium and short wavelengths, respectively. The existence of these three kinds of colour detectors makes a basis of colour perception. Cones are concentrated in the fovea, a small area near the center of the retina. In bright light the vision is called photopic and the cone cells mediate the colour perception. [4] When the light level is low, the vision is scotopic and mediated by retinal
rod cells, which do not detect colour differences. Basically, everyone is colour blind in dim lighting.

![Spectral sensitivities of cone cells](image)

**Figure 2.2:** Spectral sensitivities of cone cells, adapted from [3]

Because the vision is based on these three cones that are preferentially sensitive to red (L-cones), green (M-cones) and blue (S-cones) wavelengths, the human vision is called trichromatic. Colour perception done by the cones can be studied by the colour-matching experiment, which has been more broadly presented in [21]. The opponent colour theory states that since there is some overlap in the wavelengths to which the different cones react, it is efficient to record the differences between the responses of cones [3]. Notice, that the RGB model is not a straightforward presentation of the human colour vision capabilities but merely a convenient way of representing colours.

### 2.2.5 Perception of motion

Movements of objects can be divided to rigid movements of objects and form changes of objects. It is suggested that the visual field is monitored by motion detectors with receptive fields extended over time and space. They analyse both the form and the motion of objects in motion [31]. The object's perceived velocity does not correspond to its absolute velocity but rather to its relation to other objects in the visual field. The most important cues to motion are change in the object's angular position and size. According to [32], the observer can analyse the direction of motion by pursuit eye movements and fixations (see Chapter 2.3). The motion is detected as transformations of spatial patterns of light entering the eye [30]. Movement of all the images in the environment relative to the viewer creates a pattern of retinal motion referred to as optic flow, which creates a perception of self-motion [8].

### 2.3 Eye movements

#### 2.3.1 Versions and vergences

There are three muscles that attach the eye to the head allowing its movement in three directions: vertically, horizontally and torsionally. Movements of both the eyes simultaneously to a same direction are called versions. There are two kinds of versions: fast discontinuous eye movements, called saccades (presented in more details in Chapter 2.3.2), and pursued movements. The latter ones are usually extremely hard to produce without an object moving across the visual field [1].
Vergence is the eyes moving to opposite directions, normally to focus on an object of interest. To look at an object that is located closer the eyes move closer to each other (convergence) and to look at an object further away the eyes move away from each other (divergence) [7].

Vergence and version movements occur almost always in concurrence. Even though vergence movements are smooth, pure vergence is almost never observed and especially divergence is always associated with saccades. According to [7], matches between large receptive fields are not the only trigger to initiate vergence movements, but also texture contours [1], [9].

2.3.2 Saccadic movements

A person actively searches for information in the surrounding scene. The eyes scan a scene by fixating to one highlight after another. These movements are called saccades and they are taken care of by the six eye muscles. During the saccades, visual image is suppressed.

The eyes are constantly moving – between short saccades there is always some jitter and instability. These involuntary fixational eye movements can be divided to microsaccades, ocular drifts and ocular microtremor. Microsaccades are saccades with the smallest amplitude whose role in vision remains unknown, ocular drifts are slow random drifting movements and microtremors are high-frequency jitter. The importance of microsaccades has been in discussion and it is not even sure still, if they only exist in unnatural situations, where the subject has to maintain prolonged fixation [1].

Even though the amplitude of the fixational eye movements is extremely small, the receptive fields of neurons may be small enough to capture these movements. In addition, other kinds of fixational eye movements also exist, such as corrective saccades and post-saccadic drifts.

2.4 Stereopsis

2.4.1 Binocular geometry

The human eyes are separated horizontally by a distance of approximately 6.3 cm on average [20], which provides both the eyes with a unique image of the world. The existence of two different retinal images is called binocular disparity [9]. The difference in retinal images provides cues about the relative distance of objects and the depth structure of the objects and their environment.

The effectiveness of binocular disparity is strongest at close distances. As distance increases the retinal disparities become smaller and the ability to discriminate differences in depth diminishes.

The area where the points in space are imaged on corresponding retinal points for a given degree of convergence is called horopter. Around the horopter, the region where the two retinal images are seen as one single image is called Panum’s area. Within the area the points on one retina will fuse with a single point on the other retina. The points further away produce double images with uncrossed disparity and the points closer to the viewer produce double images with crossed disparity. Outside the range of Panum’s area depth can still be perceived, but stimuli of the two eyes cannot be fused. Stereoscopic sensitivity is remarkably good around horopter [3], [7], [10], [11], [12].
2.4.2 Binocular stereopsis

The term *binocular vision* is used for a situation with a large area of binocular overlapping used to obtain cues for coding depth [18]. Because of the eyes’ vantage points are horizontally separated, they get different images of the world simultaneously, and this is called binocular stereopsis or just stereopsis. Even though one eye can also give cues for coding depth, such as accommodation, occlusion, linear and aerial perspective, relative size, relative density and motion parallax, stereopsis enhances the ability to discriminate differences in depth. The binocular stereopsis uses the difference of two retinal images to get depth information. Binocular stereopsis is studied in [1], [7], [10], [13], [11], [17].

2.4.3 Binocular depth cues

Human ability to estimate distances is quite good when there are many sources of information available [22]. According to [23] and [24], the most important depth cues seem to be binocular stereopsis (see Sub-section 2.4.2) and motion parallax.

Motion parallax cues are created when an observer moves his eyes or head. The fixation point stays automatically on a specific point and the closer and further images move relatively to each other on the observer’s retina. There are different kinds of motion parallax depth cues. Head movements facilitate the depth estimation, but also eyeball movements have the same kind of effects. Artefacts in the temporal domain (e.g. motion blur, display persistence) will affect the motion parallax depth cues [14].

Binocular depth cues are a consequence of both eyes observing the scene at slightly different angles. The mechanism of binocular depth estimation has two parts – vergence and stereopsis. Vergence is the process, in which both eyes take a position which minimizes the difference of the visual information projected in both retinae. The angle between the eyes is used as a depth
cue. With the eyes converged on a point, stereopsis is the process which uses the residual disparity of the surrounding area for depth estimation relative to the point of convergence. Binocular depth cues are the ones most often associated with “3D cinema”. However, binocular vision is quite vulnerable to artefacts – lots of factors can lead to an “unnatural” stereo-pair being presented to the eyes. As HVS is not prepared to handle such information, binocular artefacts can lead to nausea and “simulator sickness” [25]. It is worth saying, that around 5% of all people are “stereoscopically latent” and have difficulties assessing binocular depth cues [3], [26]. Such people have a perfect depth perception, only they rely mostly on other depth cues.

Accommodation, the ability of the eye to change the optical power of its lens in order to focus on objects at various distances, explained more closely in Chapter 2.4.4.

For longer distances, binocular depth cues become less important, and HVS relies on pictorial cues for depth assessment. These are depth cues that can be perceived even with a single eye – shadows, perspective lines, texture scaling.

![Figure 2.4: Importance of depth cues as a function of the distance [14]](image)

### 2.4.4 Convergence-accommodation rivalry

Accommodation is a process in which the eye’s lens changes its optical power in order to bring an object to focus on the retina. Vergence is defined by the movement of the two eyes in opposite directions, mainly to gaze at an object of interest with both the eyes.

Accommodation and vergence work normally in concerted action, because normally the objects which are in focus are also fixated upon. As the lens accommodates to fixate to an object it is quite normal for the eyes to have vergence movements. Also, eyes converging to bring an object to focus of interest, accommodation may be produced. The convergence movements occur in response to crossed retinal disparity, whereas divergence movements occur for uncrossed disparities [28] (see also Sub-section 2.4.1). In a stereoscopic display this might introduce problems since the eyes converge to focus on an object but the lens accommodation stays on the screen where the image is sharpest [20], [27].
3 Modelling of human visual system

3.1 HVS modelling overview

One common reason for developing HVS models is to improve our ability to make high quality reproductions of the visual world. There are two general approaches in visual quality estimation: feature-based and model-based quality estimation. The feature-based approach, where an algorithm tries to assess the presence of visual artefacts is discussed in the next section. The model-based approach tries to emulate the visual process, and to estimate the perceptibility of visual details - and thus - the impact that changes to an image could have on the visual quality.

There are two types of HVS models - neurobiological ones and models based on psychophysical properties of the vision. Models based on neurobiology aim to estimate the actual low-level process in the eye and optical nerve. However, these are not useful in real-world application, because of their overwhelming complexity [33].

The psychophysical models are used to predict aspects of the human vision, which are relevant to picture quality, such as colour perception, contrast sensitivity, temporal and pattern masking. Such models are built upon psychophysical experiments, and are typically implemented as a chain of processing blocks, as shown in Figure 3.1.

The first stage models the non-linear sensitivity of HVS to light, also known as luminance masking or lightness non-linearity [34]. Then the colour perception is modelled by transforming the input signal into an adequate colour space. In HVS both luminance masking and colour processing happen simultaneously as part of the same process. The non-linearity of lightness perception is included in some colour spaces such as CIE Lab. However, some early visual models omit the colour processing stage for simplicity.

Measurements of receptive fields in the primary visual cortex show that HVS operates on multiple channels, each one tuned to different spatial frequency and orientation [35]. This can be modelled by a multi-resolution filter bank [36] or wavelet decomposition [37]. Additionally, it is believed that there are channels tuned to different temporal frequencies. This includes one temporal low pass, transient channel, and one or two band-pass, transient channels [38][39].

Once the visual information is decomposed into channels, local contrast mechanisms come into effect. The HVS response depends predominantly on local variations of luminance (i.e. contrast) and very little on the absolute luminance value. This HVS property is known as Weber-Fechner law [34]. However, the contrast perception also depends on local image content, and adaptation to specific colour or luminance model, which makes the precise modelling much more complex [40][41][42].

The sensitivity of HVS varies for different spatial frequencies. This is typically modelled by Contrast Sensitivity Function (CSF). For simplicity, colour and pattern sensitivity are assumed to be independent, and are modelled independently. Achromatic CSF models are described in [43],
and colour CSF measurements are described in [44]. The CSF also depends in the temporal frequency of the stimuli, an effect which is modelled as spatio-temporal contrast sensitivity functions. Such functions are described in [45], [46], [47].

Masking and facilitation are the two sides of the same effect. Masking occurs, when a feature which is visible by itself cannot be detected due to the presence of another visual stimulus. Facilitation is the opposite — a feature which is not visible by itself becomes visible due to the presence of another stimulus. Different types of spatial masking effects are described in [48]. Temporal masking effects, where visibility thresholds change due to temporal discontinuity before or after a visual event, are described in [49].

### 3.2 Single and multichannel models

The early models of HVS were done as a single spatial filter, following the properties of contrast sensitivity function. In this approach, the ability to detect certain visual feature is measured on the filtered version of the image, using a threshold criterion. Such models were developed for achromatic images [50] and colour ones [51]. Later, single-channel models evolved to include spatiotemporal CSF models [52]. The current state-of-the-art in single channel psychophysical HVS modelling is ST-CIELAB (spatiotemporal CIELAB), a modified CIE Lab colour-space where perceptibility of stimuli is measured as distance of that space. ST-CIELAB converts RGB data to CIE Lab format, with additional pre-processing stage which is based on spatial, temporal and chromatic model of the CSF. ST-CIELAB is described in [53].

Multi-channel HVS models first decompose the input data into many sub-bands with different spatial frequency and orientation. Detectability of visual stimuli is measured independently for each channel using different threshold criteria. An example of quality metrics, which use multi-channel HVS models are Visual Difference Predictor (VDP) [54], and Sarnoff JND [55]. VDP produces a visibility map, which shows where the two images differ in perceptual sense. Sarnoff JND is a proprietary visual quality metric for colour video. A more sophisticated spatio-temporal CSF model, which also takes into account pattern masking is used in the Perceptual Video Distortion (PDM) metric presented by S. Winkler in [56].

### 3.3 Models of HVS properties

#### 3.3.1 Luminance

In psychophysics, luminance is a measure for the “apparent intensity”. The HVS sensitivity to luminance is measured for different wavelengths using subjective tests where users match light intensity of different light sources. Additionally, the term luminance is used for the achromatic channel of an image.

CIE defines luminance with the function which defines the ratio of radiance between two wavelengths, which have the same apparent intensity when using direct comparison [57]. One wavelength is a reference one with . Luminance is expressed in candelas per square meter. The function is in the CIE standard to predict if two spectral power distributions would match:

\[
\text{(3.1)}
\]

However, cannot predict the brightness matches accurately in all cases. One counter example is the Helmoltz-Kohlrausch effect, where chromatic stimuli having the same luminance at a white reference light, appear brighter than the reference [57]. In order to increase the prediction precision and make psychophysical experiments more consistent between users, two test methods were adopted - heteromatic flicker photometry (HFP) and minimally distinct border
In HFP, the two light sources (test and reference) alternate with a frequency between 10 and 15 Hz, and the test subject has to adjust the brightness of the test stimuli in order to minimize flicker. In MDB, both stimuli have the same wavelength, and are projected simultaneously. The user adjusts the brightness of one stimuli till the border between the two lights is indistinguishable. Currently, the aim is to develop modified CIE function, with brightness prediction which will have aditivity and proportionality properties, aditivity and proportionality properties of brightness prediction, by developing functions, and is also in better agreement with HFP and MDB test results.

Another HVS property related with luminance perception is described by Weber's law – the intensity required to detect a detail in an image increases with the intensity of the background. The Weber's law states that the ratio between the intensity of a just-noticeable stimuli and the background intensity level is a constant — , where is the just noticeable difference (JND) above the background level . Weber's law applies to other human sensory systems as well. Weber's law does not cover the full range of intensities, for which HVS operates. However, for typical outdoor lighting (above 100cd/m²), the Weber's law follows the HVS sensitivity fairly well.

Following the Weber's law, one can build a scale of logarithmically increasing sensation, one JND step at a time. The logarithmic relationship between stimulus intensity and associated sensation is known as Fechner's law. Various logarithmic functions have been applied for modelling the luminance sensitivity, as described in [57], [59].

### 3.3.2 Colour

Colour is psychophysical property of HVS, rather than a physical property of an object. Perception of colour is created by the spectral radiance of a source, its reflectance of an object, and the product of the remaining light with the spectral sensitivity of the three cone receptor types in the human eye.

Colour perception is studied by colour-matching experiments [60], where the observer is asked to adjust the intensities of primary lights to match the colour appearance of a test light. In general, observers are able to find a match using only three primary colours. The HVS feature that each colour can be represented as a mixture of three primary colours is called trichromacy. Because of such representation, there are light sources with different spectral content which produce the same colour sensation for the observer (also known as metamers).

Grassmann's law states that colour matching satisfies homogeneity and superposition [57]. Following that, the colour perception can be analysed as a linear system. Every colour can be represented using three independent values (tristimulus coordinates) and mapped onto a 3D space.

Not all combinations of three primary colours are equally suitable for such representation. The opponent colour theory [61] states that some colour can coexist in a single colour sensation, while others cannot. The results of so called hue-cancellation experiments suggest that in LGN the signal from some cones suppress the signal from other ones. For example, excitation of M-cones suppresses the signal from L-cones. This process creates opponent colours, where colour sensations are encoded as difference between cone signals. As a result, some tristimulus coordinates would represent HVS more closely than others. The exact spectre of the opponent colour pairs is still being disputed, but the general consensus is that white-black, red-green and blue-yellow components allow close representation of HVS colour perception [62].

There are various colour models, based on different set of three primary colours, each one optimized for different application. Different colour models are described in [57], [63]. CIELAB is a perceptual colour space, which is created for calculating the perceived colour difference.
between two colours [64]. Based on CIELAB colour space, additional colour models were proposed, such as S-CIELAB [65], which takes into account the spatial contrast sensitivity, ST-CIELAB [53], modelling spatio-temporal masking, and CIECAM02 [66], which models the light and colour adaptation to different surroundings.

3.3.3 Contrast sensitivity function

HVS has different sensitivity to different spatial frequencies. Visual acuity, optical parameters of the eye and the multichannel processing determine the Contrast Sensitivity Function (CSF) of HVS.

The Modulation Transfer Function (MTF) of the eye determines the visual "resolution" and limits the maximum frequency throughput of HVS. The MTF of the eye is estimated to fall to 1% at about 60 cycles per degree, for standard lighting conditions and pupil size of 2mm [67]. Additionally, the maximum density of cones on the retina is around 32 arc min, which corresponds to sampling frequency of 60 cpd as well. In other words, the MTF of the eye also serves as an antialiasing filter for the photoreceptors. Due to the different density of L-, M- and S-cones on the retina, visual acuity varies with the colour. The chromatic dependant spatial sensitivity of the eye (also known as chromatic CSF) has been studied by psychophysical experiments in [62] and modelled in S-CIELAB [65].

![Contrast sensitivity function for different luminance levels](image)

Figure 3.2, Contrast sensitivity function for different luminance levels, adapted from [3]

However, the CSF does not decrease monotonically with increasing of the spatial frequency. There is an optimal sensitivity between the lowest and highest visible frequencies, which results in a peak of the CSF at approximately 10 cpd. The approximate CSF for different mean luminance values is shown in Figure 3.2. The exact shape of CSF varies with luminance, colour and spatial orientation and other factors. Different CSF models are studied in [68].
Based on results of psychophysical experiments research, it is believed that HVS processes visual information in multiple parallel channels, each one sensitive to different orientation and spatial frequency [4]. It is possible CSF to be modelled as an envelope of many channels with narrower frequency selectivity [69]. A models with 6 spatial frequency channels and 8 orientations have been proposed [4], [70] and have been shown to model the psychophysical data with sufficient precision [71].

3.3.4 Temporal and spatial masking

The contrast sensitivity changes when many signals with similar properties appear simultaneously. This is studied in psychophysical experiments, where a Gabor pattern with a fixed contrast (known as mask) is superimposed onto another Gabor pattern with varying contrast (known as signal). The contrast, at which the signal can be detected, is measured versus the contrast of the mask. The resulting function is known as threshold-versus-contrast (TvC). Usually, the more pronounced is the mask, the harder is to detect the signal. However, the opposite is also possible – presence of mask to decrease the detection threshold contrast of the signal – making it easier to be seen. This effect is known as signal facilitation and usually happens when the mask and the signal have very similar properties – frequency, orientation and colour [21]. Typical shapes of TvC functions for masking and facilitation are shown in Figure 3.3.

![](Figure 3.3 Typical masking and facilitation curves. Adapted from [21])

Masking between stimuli with the same orientation is studied in [72]. Weaker masking effects also occur between signals with different orientation [73], and different colour [74] and between coloured and achromatic ones [75]. The more different are the signals in terms of spatial frequency and orientation, the less pronounced is the masking. The characteristics of CvT functions for different frequencies and orientations have a characteristic band-limited tuning response which supports the multiresolution theory [76].

Additionally, there is evidence that there is interaction between channels with different spatial and frequency characteristics, which are far apart in their tuning characteristics [4]. For example, the sum of two signals with different spatial orientation might produce sensation for frequency components with orientation that is not present in any or the original stimuli.

Temporal discontinuities in intensity also can change the visibility thresholds. Initially, the temporal sensitivity of the eye was studied in terms of critical flicker frequency (CFF) [77]. CFF is measured in Hz and represents the temporal frequency at a flickering stimulus is perceived as a
steady one. The typical experiment for measuring CFF uses Gabor pattern where the contrast is temporally modulated to reverse sinusoidally over time. Experiments for variety of spatial and temporal frequencies showed that with increasing the temporal frequency, the spatial CSF changes its shape from band-pass to low-pass. Such behaviour suggests possible multichannel organisation of the temporal visual mechanisms similar to the spatial ones [78].

A visual model, which takes into account the spatial masking, has been presented in [54]. The model tries to predict the visibility of artefacts by calculation the visual threshold map over an image. The iCAM model, presented in [79] also includes spatial masking, but instead of predicting threshold differences, aims at predicting image quality well above threshold.

ST-CIALAB uses a single-channel spatiotemporal model, which predicts the perceptual differences between a reference and test streams. Another temporally-aware model is used in DVQ [80] aimed at assessing detection probability of threshold differences in video. The PDM metric [21] models the spatiotemporal CFS using a chain of two processing stages – the first stage uses two filters to model the transient and stationary channels in the temporal CSF, and steerable pyramid decomposition (adapted from [81]) to model the multi-channel cortex transform, responsible for spatial CSF.

### 3.4 Modelling binocular vision

Modelling of the binocular vision has been the starting point for many algorithms, which aim to extract scene depth by using two observations of a scene. In their theory Marr and Poggio present a model based on zero crossings [82]. Mayhew and Frisby [83] showed that the experienced percept could not be satisfyingly explained simply by considering zero crossings. They therefore suggested that the maximum and minimum values in the convoluted images should be matched as well. However, the majority of binocular vision algorithms are optimized for machine vision use [84]. Modelling of the binocular vision in perceptually correct way is still a challenging task, which requires that one takes into account the specific way the HVS handles the stereoscopic visual input.

The eye separation causes two different depth cues: vergence and stereopsis. The nervous system has two types of information about eye position – one based on the image falling on the retina and another based on the signal coming from motor pathways, stating the movement of eye muscles. Movements of the eyes cause vergence, a depth cue given by the angle between the two eyes when they are fixated to an object. However, studies by Foley [85] have indicated that vergence has minor impact on depth perception.

The difference between the images, captured by each retina is called binocular disparity and produces stereopsis [83],[86]. The disparity is used by the human brain to calculate the relative depth of the images. Binocular fusion and stereopsis are not synonymous: stereopsis can occur even when horizontal disparities are too large to be fused. Also, one may experience fusion of disparate images without a sense of stereoscopic depth [87].

If there are no matching images and the stereoscopic image cannot be fused from the two separate images, there might be an attempt to match the images by vergence movements of the eyes. If this still does not yield a single stereoscopic image, the non-matching features have dominance and suppression in reciprocal periods. When two stimuli with highly different contrast but similar orientation are presented dichoptically, only the high contrast stimulus is perceived. Berardi et al. [88] say that when high and low contrast gradients are presented simultaneously, one to each eye, the cell’s response to the low contrast stimulus is suppressed. According to Blake [87], this phenomenon called binocular rivalry only happens when fusion is not possible. The rivalry occurs on many levels of processing throughout the visual system [87] and can make the visual processing unstable and unpredictable.
A theory contrary to the one introduced by Blake is that of Wolfe [89] stating that there are two parallel pathways, one mediating stereopsis and another mediating binocular rivalry. The rivalry pathway according to Wolfe is active all the time. This conflicts with the proposition stating that the rivalry only occurs if fusion is not possible and is thus optional.

In order to define the relative distance between two distant objects the HVS must register the distance to the objects as well as their disparities. This process is called the scaling of disparity information [90] and is the main reason for the effectiveness of binocular disparity being the strongest at close distances. According to Patterson et al., the maximum effectiveness is attained within a distance of up to two meters [90]. As the distance from the object increases beyond this, the retinal disparities become smaller and the ability of the HVS to discriminate differences in depth diminishes.

At longer distances, the binocular cues have less importance and pictorial cues such as shadows, perspective lines, textures, size, and colours are more important. Failure to present credible pictorial cues might cause effects such as puppet theatre effect or cardboard effect [14] and thus destroy the sensation of depth. Zimmerman et al. [91] studied the human ability to discriminate depth from pictorial cues. According to them, pictorial cues are used in connecting pieces of surfaces together to create a coherent entity. They noticed that the ability to make accurate depth estimations is reduced when the surfaces are disconnected, and also that perception of depth using perspective lines is accurate only in local areas.

According to Nawrot [107] and Ohtsuka [108], the most important depth cues seem to be binocular stereopsis and motion parallax. Artefacts in temporal domain, such as motion blur or display persistence, can affect motion parallax depth cues [14]. The function of motion parallax as a depth cue has been studied by Rogers, Graham, and Wallach [68], [92], [93]. Johnston et al. [94], found that when the distance increases the weight of the texture increases. They speculate that this might be because of the decreasing importance of stereopsis on longer distances. According to Frisby et al. [95], stereo perception is not calibrated by the texture of the object when combined with stereo viewing.

Machine vision algorithms, which employ feature based cooperative stereo-matching, are found to have performance very similar to the stereoscopic HVS [110]. A sub-class of cooperative stereo algorithms, namely coherence-based stereovision ones, are specifically designed with modelling the human vision in mind [111]. And improved version of the coherence-based stereo matching algorithm, proposed in [112] is based on biologically plausible model which simulates the human gaze, and can produce disparity map of a scene with very few iterations.

Coherence-based stereo algorithms share both the advantages (fast and noise-insensitive) and weaknesses (inexact and context sensitive) with the stereoscopic HVS [112]. Such algorithms produce two data structures simultaneously: one is an observation of the scene from a virtual, straight-ahead perspective, which directly corresponds to the “cyclopean view” experienced by human observers; the other is a relative disparity map, which corresponds to the binocular disparity depth cues that produce stereopsis.
4 Visual quality metrics

4.1 Model-based and feature-based approach

There are two big families of visual quality metrics—model-based and feature-based ones. Model-based metrics take the “top-down” approach, modelling the human vision, and estimating the visibility of artefacts. Feature-based metrics estimate quality in a “bottom-up” fashion, trying to assess presence of visual artefacts through signal processing methods, and statistically estimate their impact over the generally perceived quality. Both approaches have their advantages and disadvantages.

Model-based quality metrics start with methodological study of the processes in vision and cognition which occur in HVS. Then, each visual stage is modelled in a chain of signal processing blocks, starting with refraction in the lens and photoreceptors in the retina, and ending with masking effects and visual salience as result of brain cortex processes. The parameters of the visual models are tuned according to the results of psychophysical experiments. Finally, the perceptual difference between two visual stimuli (images or video) is used to estimate the visibility of artefacts in a “test image” as opposed to an artefact free “reference image”.

The advantage or a model-based visual quality metric is that if properly done, it should be generally applicable to wide range of images or video streams, and always produce meaningful results. However, model-based metrics suffer from so-called supra-threshold problem. The psychophysical experiments are usually designed to operate in the near-threshold range of vision, where the tests subject gives a binary answer if a visual stimulus is barely visible or not visible at all. The stimuli used in these experiments are simple lines, dots or crosses. How the results for threshold visibility of simple stimuli can be extrapolated for natural images in the supra-threshold visibility range remains largely unknown and up to the intuition of the psychophysicists [3].

Feature-based quality metrics work by evaluating various features in images or video streams and using a weight-based function to estimate their impact on the visual quality. Feature-based metrics are usually designed for specific usage scenarios. First the typical artefacts which might occur in the usage case are identified. Then, natural images or videos impaired with these artefacts are created. The relative impact of each artefact on the overall quality is estimated through subjective quality experiments. An algorithm which can identify the presence and amount of each of these artefacts is designed. Finally, the assessments for each artefact are combined, with different weights, which are “tuned” so the prediction of the metric statistically matches the subjective quality score for the set of test images.

The advantage of the feature-based metrics is that even if not versatile, they are more computationally effective than model-based ones. This makes feature-based metrics suitable for specific applications, where a limited set of artefacts are expected and computational efficiency is a must. Main disadvantage of feature-based metrics is that they require large test set databases and sufficient number of participants in subjective quality tests. Otherwise feature-based metrics suffer from the so-called statistical pitfall. Given a small set of subjective vs. objective pairs, and enough degrees of freedom, one can always achieve a good match of MOS vs. predicted quality. As a result, the feature-based metric becomes “over-fitted” for the test set, and performs very well over the images or videos in the test database, but fails in other cases [33]. This effect can be somehow mitigated, if the feature-based metric is designed with good understanding of the visual process, thus avoiding the introduction of unnecessary, “unnatural” degrees of freedom in the statistical model.
4.2 Metric classification

Quality metrics can be divided into three categories based on the information that is needed about the reference video. The first group of metrics are called full-reference (FR) metrics. They need the entire reference video in order to compare each frame of it to the frames of the video under test. The other far end of the metrics is no-reference (NR) metrics, which uses no reference information at all, but looks only at the video under test. The problem with NR metrics is, that the actual content might be confused with some distortions. Reduced-reference (RR) metrics uses some features from the reference video but not all of it like full-reference metrics, thus keeping the amount of information manageable.

There are several examples of full-reference metrics, such as DCTune [124], [125]. It computes the JPEG quantization matrices that achieve the maximum compression for a specified perceptual distortion given a particular image in particular viewing conditions. Another example is a picture quality assessment system based on a perceptual weighting of the coding noise [126]. An objective measurement tool for MPEG video quality has also been presented by Tan et. al. [127]. It simulates the delay and temporal smoothing effect of observer responses, the nonlinear saturation of perceived quality and the asymmetric behaviour with respect to quality changes. Hekstra et. al. [128] proposed a perceptual video quality measure that uses a linear combination of the loss of edge sharpness, the colour error and the temporal variability of the reference video. Wang et. al. [142] presented a video quality assessment method based on a structural similarity index. It is applied to colour video by computing the structural similarity index for each colour channel.

Reduced-reference information is in use in a video quality metric designed by Wolf and Pinson [129]. It extracts low-level features from spatio-temporal blocks of sequences. Another reduced-reference metric was proposed by Horita et. al. [130]. It is based on 26 low-level spatial features computed from the luminance image and the Sobel edge filters.

The majority of no-reference metrics are based on estimating blockiness. Wu and Yuen [131] introduced a NR blocking impairment metric, which measures the horizontal and vertical differences between the columns and rows at block boundaries. Baroncini and Pierotti [132] have a metric based on multiple filters to extract significant vertical and horizontal edge segments due to blockiness. Wang et. al. [133] introduce a model the blocky image as a non-blocky image interfered with a pure blocky signal. Vlachos [134] used an algorithm based on the cross correlation of subsampled images. There are also methods that use DCT coefficients directly from the encoded bit stream. Coudoux et. al. [135] introduced such a metric, combining the detection of vertical block edges with a number of sensitivity and masking models applied in the DCT domain. Gastaldo et. al. [136] presented an NR measurement approach using a neural network. There are also other types of distortions than compression artefacts, and for these types NR metrics that don't base on estimating blockiness. Object boundaries are usually represented by sharp edges, and based on this assumption Marziliano et. al. [137] proposed a blurriness metric. Winkler et. al. [138], [139] combined blockiness, blurriness and jerkiness artefact metrics for real-time NR quality assessment. Another no-reference video quality metric was presented by Caviedes and Oberti [140].

It is based on several artefacts and features, such as blocking, ringing, clipping, noise, contrast, sharpness and their interactions.
4.3 Overview of visual quality metrics

4.3.1 PSNR

PSNR is the logarithm of the inverse of Mean Square Error (MSE) between two images/video frames. MSE is the dominant quantitative performance metric in the field of signal processing. MSE exhibits weak performance and widely criticized when dealing with perceptually important signals such as speech and images [141]. However it is still used heavily since it is simple, inexpensive to compute, memory-less and is excellent metric for optimization.

Some of the properties of MSE do not correlate with the perception of visual data. MSE is independent of temporal or spatial relationships between the samples of the original signal. MSE is independent of any relationship between the original signal and the error signal. MSE is independent of the signs of the error signal samples. Also all signal samples are equally important in MSE. However two different images which have the same MSE value, can have very different perceptual quality [15], [141].

4.3.2 SSIM

Cluster of natural images occupies an extremely tiny portion in the image space and they are highly structured with samples having strong neighbour dependencies. Also human visual system is an information extractor that seeks to identify and recognize objects, highly sensitive to the structural distortions and automatically compensates for the non-structural distortions. In order to compensate the disadvantages of MSE and incorporating the importance of structural information SSIM metric is introduced [SSIM]. It is made up of easy to compute statistics (mean, variance and covariance of small patches inside a frame) of luminance comparison, contrast comparison and structural comparison as shown in Figure 4.1. It can also be implemented over multiple scales [144], or in the wavelet domain [145].

![Diagram of the structural similarity (SSIM) measurement system, adapted from [143]](image-url)

Figure 4.1. Diagram of the structural similarity (SSIM) measurement system, adapted from [143]
4.3.3 VSSIM
There are strong correlations between adjacent video frames (temporal and spatio-temporal signal structures). Also video contains perceptually important structured motion. Rather than averaging SSIM of individual frames, two adjustments are made to generate a weighted SSIM for video (VSSIM [142]). The first is based on the observation that dark regions usually do not attract fixations, therefore should be assigned smaller weighting values. Second adjustment is by assigning smaller weights for frames with large global motion since image distortions are perceived differently when the background of the video is moving very fast. Overall system architecture is given in Figure 4.2.

![Figure 4.2. Proposed video quality assessment system, adapted from [142]](image)

4.3.4 CW-SSIM
SSIM algorithm is highly sensitive to translation, scaling and rotation of images. By using complex wavelet transform domain version of SSIM (CW-SSIM), new metric becomes insensitive to “non-structured” image distortions that are typically caused by the movement of the image acquisition devices [145].

4.3.5 ICIQ
ICIQ is another quality metric, based on the concept of “structural similarity” [146]. However, while the SSIM family of metrics are based on local statistics, ICIQ uses “similarity of adaptive scales” as the key indicator of structural similarity. The adaptive scales are determined by a modification the intersection of confidence intervals (ICI) algorithm [147] originally developed for image restoration. Changes in the image structure reflect into changes of the adaptive scales on an image. ICIQ measures the differences in the adaptive scales between two images, thus merging scale and structure in one concept.

ICIQ is a compound metric based on two terms - Window Term (WT), which is a difference map of the adaptive scales for two images, and Intensity Term (IT), which measures the intensity differences between two images, normalised to their dynamic range. The WT term is sensitive to structural changes, and IT term is reflecting the structural-preserving impairments, such as mean shift and image negation. Maps of adaptive scales, WT and IT terms over the “Lena” image are shown in Figure 4.3.
Changes in the image structure are reflected into changes in the adaptive scales found for an image. Changes in the image structure are reflected into changes in the adaptive scales, and that the adaptive scales are more sensitive to structural changes than the conventional measures based on local statistics. Comparison between ICIQ and SSIM over the images from the LIVE database R1 and R2 [118] show that ICIQ gives more appropriate scores for some image artefacts (see Figure 4.4) and performs comparably over a broader set of image impairments. The results demonstrate that a quality indicator solely based on the adaptive scales (mWT), is producing satisfactory results, demonstrating that similarity of adaptive scales is essential for satisfactory modelling of human quality perception.

4.3.6 **VSNR**

VSNR [148] determines whether the distortions are below the threshold of visual detection. If distortions are below then metric gives infinite value denoting perfect visual fidelity. If distortions are above threshold, low-level visual property of perceived contrast and the mid-level visual property of global precedence is used. Multiscale wavelet decomposition is used and Euclidean distance is calculated in the distortion-contrast space. This metric has low computational complexity and low memory requirements.

4.3.7 **VQM**

VQM divides sequences into spatio-temporal blocks, and a number of features measuring the amount and orientation of activity in each of these blocks are computed from the spatial luminance gradient. The features extracted from test and reference videos are then compared using a process similar to masking [151][152].

Figure 4.3. Building elements of the ICIQ metric: a) Adaptive scales for the reference image, b) Window Term for the test image, c) Intensity Term for the test image [146]
4.3.8 Sarnoff’s JND

It is based on a visual discrimination model that simulates the responses of human spatio-temporal visual mechanisms and the perceptual magnitudes of differences in mechanism outputs between source and processed sequences. Images are filtered and down-sampled using a Gaussian pyramid operation to efficiently generate a range of spatial resolutions for subsequent filtering operations. In the normalization stage, the overall gain with a time-dependent average luminance is set, to model the visual system’s relative insensitivity to overall light level, and to represent such effects as the loss of visual sensitivity after a transition from a bright to a dark scene. After normalization, three separate contrast measures (oriented contrast, flicker contrast and chromatic contrast) are calculated followed by contrast energy masking and pooling stages [157]. Block diagram of JND is in Figure 4.5.
Figure 4.5. JND Model Architecture Overview, adapted from [157]
4.3.9 DVQ

The DVQ metric is based on the Discrete Cosine Transform [153]. It incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. It also includes primitive dynamics of light adaptation and contrast masking. Its performance is similar to Sarnoff’s JND. Block diagram of DVQ is in Figure 4.6.

Figure 4.6 Overview of DVQ processing steps, adapted from [153]
4.3.10 PDM

PDM is based on a spatio-temporal model of the human visual system. Images are converted into opponent-colors space and subjected to spatio-temporal perceptual decomposition. They undergo weighting and a contrast gain control stage. Finally, differences are combined into a distortion measure [155][156]. Block diagram of PDM is in Figure 4.7.

![Block diagram of the perceptual distortion metric](image-url)
4.3.11 PQS

Several distortion factors are computed from the images using transformation by the help of Weber–Fechner’s Law, contrast sensitivity, spatial frequency weighting and the description of perceived image disturbances. Visual masking is also applied. Regression methods are applied to combine these distortion factors into a single measure [154]. Block diagram of PQS is in Figure 4.8.

4.3.12 PSNR-HVS-M

PSNR-HVS-M is Peak Signal to Noise Ratio taking into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions [149]. The model operates with the values of DCT coefficients of 8x8 pixel block of an image. For each DCT coefficient of the block the model allows to calculate its maximal distortion that is not visible due to the between-coefficient masking. PSNR-HVS-M, takes into account the proposed model and the contrast sensitivity function. In order to evaluate the performance of the metric, the authors of PSNR-HVS-M build their own set of test images [120], and performed subjective tests, where observers sorted the images in order of their visual appearance. Over their database, PSNRHVS-M has outperformed other well-known reference based quality metrics and demonstrated high correlation with the results of subjective experiments. The block diagram of the metric is shown in Figure 4.9.
4.3.13 VIF

Visual Information Fidelity (VIF) relates signal fidelity to the amount of information that is shared between two signals using an information theoretic approach and modelling human visual system and natural image space as well [150]. It also places fundamental limits on the amount of perceptually relevant information that could be extracted from a signal. VIF index exhibits superior performance relative to all other image fidelity measurement algorithms.

4.4 Quality metric evaluation

4.4.1 VQEG studies

In 1997, the Video Quality Experts Group (VQEG) was established with the aim to standardize an objective video quality metric. Initially, VQEG proposed a process of objective metric validation, and started collection submissions of objective metrics to be included in the verification process. VQEG aimed at full-reference metrics, which can evaluate the performance of block-based encoders such as MPEG-2 and H.263 with bit-rate between 768 kbps and 36 Mbps. The objective metrics were graded in respect to their ability to predict the subjective quality estimation over a test set of 2D videos. A working group within VQEG, called Independent Lab and Selection Committee (ILCS) selected a set of video sequences. The sequences were kept confidential, until all participants submitted their final metric implementations.

In the first round of tests (also known as “Phase I”) subjective and objective tests were conducted in parallel. Subjective quality assessment of the test video set was performed in multiple laboratories, using the ITU-R recommendation BT.500 [113]. The subjective test plan is described in details in [116]. The same video set was sent to each organization, which proposed an objective metric. Additionally, the results were verified by independent laboratories, as described in [115]. A total of nine algorithms were tested.
After the first round of tests, VQEG could not recommend objective video quality metric for standardization [114]. No objective model outperformed others on all cases, and no model was able to predict subjective quality scores with sufficient accuracy. However, subjective tests were successfully completed and were reported as a valuable outcome of “Phase I”.

Between 2001 and 2003 VQEG performed second round of tests, also known as “Phase II”. Only six participants sent algorithms this time. Still unable to find a clear winner, VQEG proposed four of the algorithms as having sufficient precision in predicting the subjective quality scores [117].

4.4.2 Performance parameters

Key point to quality metrics is the quality perceived by human observers, MOS. Quality metrics can be characterized by several parameters in terms of its prediction performance, with respect to subjective ratings, presented by [33] as follows:

Accuracy is the ability of a metric to predict subjective ratings. For data sets with linear relations, accuracy is defined:

\[ \text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{\sigma} \right)^2, \]  
(4.1)

where \( \bar{y} \) and \( \bar{\hat{y}} \) are the means of the respective data sets. In a nonlinear case, a mapping function must first be applied to one of the data sets.

Monotonicity measures if changes in one variable are associated with changes in the other variable. Monotonicity can be quantified by the Spearman rank-order correlation coefficient, defined:

\[ \text{Monotonicity} = \rho = \frac{\sum_{i=1}^{n} (r_1(i) - \bar{r}_1)(r_2(i) - \bar{r}_2)}{\sqrt{\sum_{i=1}^{n} (r_1(i) - \bar{r}_1)^2 \sum_{i=1}^{n} (r_2(i) - \bar{r}_2)^2}}, \]  
(4.2)

where \( r_1 \) is the rank of \( y_1 \) and \( r_2 \) is the rank of \( y_2 \) in the ordered data sets. \( \bar{r}_1 \) and \( \bar{r}_2 \) are their respective midranks.

Consistency is a parameter to measure the number of outliers of all the data points. VQEG proposed consistency to be defined as a data point \((y_i, \hat{y}_i)\), for which the prediction error is greater than twice the standard deviation \[115]:

\[ \text{Consistency} = \frac{\text{Number of outliers}}{\text{Total number of data points}}, \]  
(4.3)

The outlier ratio is then defined as the ratio of the number of outliers and the total number of data points:

\[ \text{Outlier ratio} = \frac{\text{Number of outliers}}{\text{Total number of data points}}. \]  
(4.4)

Because the perceived visual quality is a subjective measure and thus possible to be described only statistically, the question is also how well subjects agree on the quality of a certain video. The VQEG Phase I tests presented by [114] give an indication of the limits on the performance of video quality metrics. The individual differences on the tests represent the residual variance of the optical model – the minimum variance that can be achieved. According to the tests, none of the submitted metrics achieved a prediction performance statistically equivalent to the optimal model.

4.4.3 Tests sets

In order to evaluate the performance of a image quality metric, one needs a database of test images. Such database should include large set of images where different parameters, which
are expected to affect the quality are systematically varied. Each image should be annotated with a mean opinion score (MOS), obtained through subjective experiments where sufficiently large group of users grades the quality of visual stimuli. Unfortunately, large database of 3D video streams is yet to be created, and subjective 3D quality tests with large group of users are yet to be executed. One of the aims of our project is to study the subjective quality for the more restricted case of mobile 3D video, focusing on small displays and these parameters, which would affect the quality of a 3D video broadcasting over DVB-H.

There are a few publicly accessible tests sets which can be used for evaluation of visual quality metrics:

- **LIVE image quality database** – this is a publicly available test set of images, impaired with different artefacts and annotated with MOS scores [118]. The current version of the database contains 982 images, with 203 original and 779 distorted images. The distortion types include JPEG2000 compression, JPEG compression, Gaussian noise contamination, Gaussian blur, and JPEG2000 compressed images undergoing fast fading channel distortions. The performance evaluation of SSIM was initially done on the LIVE database. The creators of the database promise a similar database of test video streams to be publicly available soon.

- **Cornel A57 database** – A57 was created as evaluation test set for the VSNR quality metric. It contains only three images, impaired by six types of distortions with varying parameters [119]. The distortions are: Quantization of the LH subbands; Additive Gaussian white noise; Baseline JPEG compression; JPEG-2000 compression (with and without Dynamic Contrast-Based Quantization); Blurring by using a Gaussian filter.

- **TID2008 database** – a test set which contains 25 reference images and 1700 distorted images (25 reference images x 17 types of distortions x 4 levels of distortions) [120]. The MOS scores were obtained by 838 observers in three different countries.

- **VQEG test sets** – the original test set of videos used in VQEG experiments were provided to the participants in the tests on magnetic tapes. VQEG database is still not freely available. Instead, VQEG group provides a list of links for different video and image quality test sets [121].

- **MOBILE3DTV database** – as part of the research of MOBILE3DTV project, we created a database of impaired stereo-videos. We selected the artefacts which are most likely to affect stereoscopic video transmitted over a DVB-H channel. The database contains three stereoscopic video streams impaired by six types of artefacts. The database is available online [122]. Additional types of artefacts with various parameters can be introduced to a given stereoscopic movie by our automated MATLAB rendering framework for stereoscopic artefacts, also available online [123]. Presently, the videos in the database are not annotated with subjective MOS scores.

### 4.4.4 Comparison of 2D quality metrics for 3D video artefacts

This section presents the test results that are generated for the evaluation of 3D video artifacts on stereo video. Impaired video database that is used in the tests and detailed information about the creation of impaired videos are available in the Mobile3DTV Project website [122]. Mainly two different set of visual quality metric tools are used for objective distortion evaluation. First visual quality metric tool, MeTriX MuX Visual Quality Assessment Package [159] is provided by Cornell Visual Communications Lab and implements various quality metric algorithms using Matlab. This tool is aimed for image quality assessment. In the tests metrics are applied to each frame of the video and the average is taken. Results of this set of quality metric evaluations for the previously mentioned impaired-videos are presented in Table 4-1. Second visual quality metric tool, MSU VQMT, is an executable program provided by MSU Graphics & Media Lab
[160]. This tool directly takes the original and distorted videos and provides the result by a similar averaging operation on the individually run frame based computations. Detailed information about the quality metrics can be found in the group’s webpage [160]. Similarly, results of second set of quality metric evaluations are presented in Table 4-2. Moreover in Table 4-3 average time elapsed for the computation of algorithms run for one pair of impaired-videos are provided in order to gain an intuition on the complexity of the metrics. Since first tool is implemented in Matlab, we have given the computation complexity of PSNR in the other tool as well for comparison.
Table 4-1 Results of the distortion evaluations of various quality metric algorithms for the impaired video database using first tool

<table>
<thead>
<tr>
<th></th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
<th>MSSIM</th>
<th>VSNR</th>
<th>VIF</th>
<th>VIFP</th>
<th>UQI</th>
<th>IFC</th>
<th>NQM</th>
<th>WSNR</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Colour bleeding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEFT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colour bleeding</td>
<td>8.9062</td>
<td>41.5998</td>
<td>0.9908</td>
<td>0.9957</td>
<td>34.2526</td>
<td>0.9488</td>
<td>0.9359</td>
<td>0.9016</td>
<td>18.8742</td>
<td>23.8434</td>
<td>30.2123</td>
<td>29.0953</td>
</tr>
<tr>
<td>Blocking q=10</td>
<td>45.7392</td>
<td>32.6620</td>
<td>0.8527</td>
<td>0.9577</td>
<td>25.5333</td>
<td>0.3668</td>
<td>0.3724</td>
<td>0.5891</td>
<td>1.5466</td>
<td>20.8263</td>
<td>28.4009</td>
<td>20.0674</td>
</tr>
<tr>
<td>Blocking q=15</td>
<td>33.7506</td>
<td>33.9390</td>
<td>0.8880</td>
<td>0.9736</td>
<td>27.8649</td>
<td>0.4746</td>
<td>0.4554</td>
<td>0.6466</td>
<td>2.0702</td>
<td>23.6560</td>
<td>30.0377</td>
<td>21.3825</td>
</tr>
<tr>
<td>Blocking q=50</td>
<td>13.3475</td>
<td>38.4892</td>
<td>0.9521</td>
<td>0.9930</td>
<td>37.4010</td>
<td>0.7661</td>
<td>0.6334</td>
<td>0.7901</td>
<td>4.1038</td>
<td>31.1703</td>
<td>39.2779</td>
<td>25.8904</td>
</tr>
<tr>
<td><strong>Colour Mismatch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colour Mismatch</td>
<td>0.0000</td>
<td>Infinity</td>
<td>1.0000</td>
<td>1.0000</td>
<td>53.3365</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>62.6430</td>
<td>Infinity</td>
<td>Infinity</td>
<td>Infinity</td>
</tr>
<tr>
<td>CROSSTALK k=0.2</td>
<td>6.0346</td>
<td>40.8651</td>
<td>0.9868</td>
<td>0.9953</td>
<td>32.3085</td>
<td>0.8629</td>
<td>0.8074</td>
<td>0.9499</td>
<td>6.1946</td>
<td>27.0970</td>
<td>33.4136</td>
<td>28.1992</td>
</tr>
<tr>
<td>CROSSTALK k=0.6</td>
<td>52.9197</td>
<td>31.4398</td>
<td>0.8924</td>
<td>0.9580</td>
<td>20.6823</td>
<td>0.4754</td>
<td>0.4556</td>
<td>0.7286</td>
<td>2.5583</td>
<td>17.0676</td>
<td>23.7978</td>
<td>18.6893</td>
</tr>
<tr>
<td>BlockEdge s=15</td>
<td>35.2476</td>
<td>32.8159</td>
<td>0.8360</td>
<td>0.9125</td>
<td>23.6368</td>
<td>0.3520</td>
<td>0.3528</td>
<td>0.5062</td>
<td>2.5881</td>
<td>17.0200</td>
<td>24.8625</td>
<td>20.3390</td>
</tr>
<tr>
<td>BlockEdge s=30</td>
<td>134.6528</td>
<td>27.0120</td>
<td>0.6432</td>
<td>0.7760</td>
<td>17.6348</td>
<td>0.1832</td>
<td>0.1874</td>
<td>0.3950</td>
<td>1.5164</td>
<td>11.1564</td>
<td>19.1929</td>
<td>14.7574</td>
</tr>
<tr>
<td>Packet loss</td>
<td>52.3363</td>
<td>36.1396</td>
<td>0.9470</td>
<td>0.9630</td>
<td>28.7283</td>
<td>0.7586</td>
<td>0.6938</td>
<td>0.7734</td>
<td>4.4253</td>
<td>21.7036</td>
<td>29.5342</td>
<td>23.5259</td>
</tr>
<tr>
<td><strong>Colour bleeding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIGHT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colour bleeding</td>
<td>9.1705</td>
<td>41.1800</td>
<td>0.9905</td>
<td>0.9955</td>
<td>33.3615</td>
<td>0.9471</td>
<td>0.9337</td>
<td>0.9059</td>
<td>18.2547</td>
<td>23.2407</td>
<td>29.5606</td>
<td>28.5013</td>
</tr>
<tr>
<td>Blocking q=10</td>
<td>47.1234</td>
<td>32.4837</td>
<td>0.8460</td>
<td>0.9558</td>
<td>25.4176</td>
<td>0.3633</td>
<td>0.3664</td>
<td>0.5823</td>
<td>1.5266</td>
<td>20.6012</td>
<td>28.1365</td>
<td>19.7049</td>
</tr>
<tr>
<td>Blocking q=15</td>
<td>35.0541</td>
<td>33.7399</td>
<td>0.8822</td>
<td>0.9725</td>
<td>27.7707</td>
<td>0.4714</td>
<td>0.4492</td>
<td>0.6410</td>
<td>2.0522</td>
<td>23.4952</td>
<td>29.8450</td>
<td>21.0049</td>
</tr>
<tr>
<td>Blocking q=50</td>
<td>14.3402</td>
<td>38.1670</td>
<td>0.9483</td>
<td>0.9926</td>
<td>37.2157</td>
<td>0.7634</td>
<td>0.6243</td>
<td>0.7892</td>
<td>4.1044</td>
<td>30.9619</td>
<td>38.8963</td>
<td>25.3926</td>
</tr>
<tr>
<td><strong>Colour Mismatch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colour Mismatch</td>
<td>4.2248</td>
<td>43.8377</td>
<td>0.9954</td>
<td>0.9967</td>
<td>36.4806</td>
<td>0.9603</td>
<td>0.9444</td>
<td>0.9765</td>
<td>10.2737</td>
<td>26.8815</td>
<td>32.3055</td>
<td>30.8948</td>
</tr>
<tr>
<td>CROSSTALK k=0.2</td>
<td>6.0856</td>
<td>40.8225</td>
<td>0.9868</td>
<td>0.9953</td>
<td>32.5931</td>
<td>0.8710</td>
<td>0.8137</td>
<td>0.9584</td>
<td>6.3847</td>
<td>26.8573</td>
<td>33.1733</td>
<td>28.0498</td>
</tr>
<tr>
<td>CROSSTALK k=0.6</td>
<td>53.0590</td>
<td>31.4266</td>
<td>0.8930</td>
<td>0.9578</td>
<td>21.4934</td>
<td>0.4954</td>
<td>0.4669</td>
<td>0.7433</td>
<td>2.6215</td>
<td>17.0584</td>
<td>23.7921</td>
<td>18.7117</td>
</tr>
<tr>
<td>BlockEdge</td>
<td>35.0034</td>
<td>32.8370</td>
<td>0.8375</td>
<td>0.9133</td>
<td>23.6586</td>
<td>0.3564</td>
<td>0.3557</td>
<td>0.5139</td>
<td>2.5963</td>
<td>16.9114</td>
<td>24.7120</td>
<td>20.1844</td>
</tr>
<tr>
<td>s=15</td>
<td>BlockEdge s=30</td>
<td>Packet loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>133.2813</td>
<td>27.0454</td>
<td>0.6459</td>
<td>0.7773</td>
<td>17.6800</td>
<td>0.1849</td>
<td>0.1881</td>
<td>0.4003</td>
<td>1.5165</td>
<td>11.0552</td>
<td>19.0523</td>
<td>14.6174</td>
<td></td>
</tr>
<tr>
<td>58.7963</td>
<td>34.8059</td>
<td>0.9404</td>
<td>0.9597</td>
<td>26.6315</td>
<td>0.7300</td>
<td>0.6684</td>
<td>0.7642</td>
<td>4.1807</td>
<td>19.6564</td>
<td>27.2483</td>
<td>22.0522</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colour bleeding</th>
<th>JOINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0384</td>
<td>41.3899</td>
</tr>
<tr>
<td>46.4313</td>
<td>32.5729</td>
</tr>
<tr>
<td>34.4023</td>
<td>33.8394</td>
</tr>
<tr>
<td>13.8438</td>
<td>38.3281</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colour Mismatch</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
<th>Inf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1124</td>
<td>Infinity</td>
<td>0.9977</td>
<td>0.9984</td>
<td>44.9085</td>
<td>0.9802</td>
<td>0.9722</td>
<td>0.9882</td>
<td>36.4583</td>
<td>Infinity</td>
<td>Infinity</td>
<td>Infinity</td>
<td>Infinity</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crosstalk k=0.2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0601</td>
<td>40.8438</td>
<td>0.9868</td>
<td>0.9953</td>
<td>32.4508</td>
<td>0.8670</td>
<td>0.8105</td>
<td>0.9541</td>
<td>6.2897</td>
<td>26.9772</td>
<td>33.2934</td>
<td>28.1245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52.9894</td>
<td>31.4332</td>
<td>0.8927</td>
<td>0.9579</td>
<td>21.0879</td>
<td>0.4854</td>
<td>0.4612</td>
<td>0.7360</td>
<td>2.5899</td>
<td>17.0630</td>
<td>23.7949</td>
<td>18.7005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.1255</td>
<td>32.8264</td>
<td>0.8367</td>
<td>0.9129</td>
<td>23.6477</td>
<td>0.3542</td>
<td>0.3542</td>
<td>0.5100</td>
<td>2.5922</td>
<td>16.9657</td>
<td>24.7873</td>
<td>20.2617</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BlockEdge s=15</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>133.9671</td>
<td>27.0287</td>
<td>0.6445</td>
<td>0.7766</td>
<td>17.6574</td>
<td>0.1841</td>
<td>0.1877</td>
<td>0.3977</td>
<td>1.5164</td>
<td>11.1058</td>
<td>19.1226</td>
<td>14.6874</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55.5663</td>
<td>35.4727</td>
<td>0.9437</td>
<td>0.9614</td>
<td>27.6799</td>
<td>0.7443</td>
<td>0.6811</td>
<td>0.7688</td>
<td>4.3030</td>
<td>20.6800</td>
<td>28.3912</td>
<td>22.7890</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4-2 Results of the distortion evaluations of various quality metric algorithms for the impaired video database using second tool

<table>
<thead>
<tr>
<th>VQM</th>
<th>LEFT</th>
<th>RIGHT</th>
<th>JOINT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VQM</td>
<td>Blocking Beta</td>
<td>Blocking Beta</td>
</tr>
<tr>
<td></td>
<td>original</td>
<td>distorted</td>
<td>original</td>
</tr>
<tr>
<td>Colour bleeding</td>
<td>1.0261</td>
<td>10.9045</td>
<td>11.1807</td>
</tr>
<tr>
<td>Blocking q=10</td>
<td>2.1911</td>
<td>10.9045</td>
<td>63.9590</td>
</tr>
<tr>
<td>Blocking q=15</td>
<td>1.9228</td>
<td>10.9045</td>
<td>35.9060</td>
</tr>
<tr>
<td>Blocking q=50</td>
<td>1.1075</td>
<td>10.9045</td>
<td>15.1758</td>
</tr>
<tr>
<td>Colour Mismatch</td>
<td>0.0000</td>
<td>10.9045</td>
<td>10.9045</td>
</tr>
<tr>
<td>Crosstalk k=0.2</td>
<td>0.9291</td>
<td>10.9045</td>
<td>10.9952</td>
</tr>
<tr>
<td>Crosstalk k=0.6</td>
<td>2.6987</td>
<td>10.9045</td>
<td>10.9358</td>
</tr>
<tr>
<td>BlockEdge s=15</td>
<td>1.6623</td>
<td>10.9045</td>
<td>37.8161</td>
</tr>
<tr>
<td>BlockEdge s=30</td>
<td>3.1367</td>
<td>10.9045</td>
<td>63.4908</td>
</tr>
<tr>
<td>Packet loss</td>
<td>2.4951</td>
<td>10.9045</td>
<td>12.0582</td>
</tr>
<tr>
<td>Colour bleeding</td>
<td>1.0724</td>
<td>10.6524</td>
<td>10.9930</td>
</tr>
<tr>
<td>Blocking q=10</td>
<td>2.2623</td>
<td>10.6524</td>
<td>63.8741</td>
</tr>
<tr>
<td>Blocking q=15</td>
<td>1.9886</td>
<td>10.6524</td>
<td>35.1643</td>
</tr>
<tr>
<td>Blocking q=50</td>
<td>1.1602</td>
<td>10.6524</td>
<td>14.9869</td>
</tr>
<tr>
<td>Colour Mismatch</td>
<td>0.6650</td>
<td>10.6524</td>
<td>10.7388</td>
</tr>
<tr>
<td>Crosstalk k=0.2</td>
<td>0.9931</td>
<td>10.6524</td>
<td>10.7820</td>
</tr>
<tr>
<td>Crosstalk k=0.6</td>
<td>2.7930</td>
<td>10.6524</td>
<td>10.9710</td>
</tr>
<tr>
<td>BlockEdge s=15</td>
<td>1.6708</td>
<td>10.6524</td>
<td>37.6599</td>
</tr>
<tr>
<td>BlockEdge s=30</td>
<td>3.1491</td>
<td>10.6524</td>
<td>63.5233</td>
</tr>
<tr>
<td>Packet loss</td>
<td>2.7215</td>
<td>10.6524</td>
<td>11.8380</td>
</tr>
<tr>
<td>Colour bleeding</td>
<td>1.0492</td>
<td>10.7784</td>
<td>11.0868</td>
</tr>
<tr>
<td>Blocking q=10</td>
<td>2.2267</td>
<td>10.7784</td>
<td>63.9166</td>
</tr>
<tr>
<td>Blocking q=15</td>
<td>1.9557</td>
<td>10.7784</td>
<td>35.5352</td>
</tr>
<tr>
<td>Blocking q=50</td>
<td>1.1339</td>
<td>10.7784</td>
<td>15.0814</td>
</tr>
<tr>
<td>Colour Mismatch</td>
<td>0.3325</td>
<td>10.7784</td>
<td>10.8216</td>
</tr>
<tr>
<td>Crosstalk k=0.2</td>
<td>0.9611</td>
<td>10.7784</td>
<td>10.8886</td>
</tr>
<tr>
<td>Crosstalk k=0.6</td>
<td>2.7459</td>
<td>10.7784</td>
<td>10.9534</td>
</tr>
<tr>
<td>BlockEdge s=15</td>
<td>1.6665</td>
<td>10.7784</td>
<td>37.7380</td>
</tr>
<tr>
<td>BlockEdge s=30</td>
<td>3.1429</td>
<td>10.7784</td>
<td>63.5070</td>
</tr>
<tr>
<td>Packet loss</td>
<td>2.6083</td>
<td>10.7784</td>
<td>11.9481</td>
</tr>
</tbody>
</table>
Table 4-3 Performance information of the visual quality metrics of the tools

<table>
<thead>
<tr>
<th>Quality Metric</th>
<th>Average Elapsed Time per frame (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>42.10</td>
</tr>
<tr>
<td>SNR</td>
<td>43.89</td>
</tr>
<tr>
<td>PSNR</td>
<td>44.04</td>
</tr>
<tr>
<td>SSIM</td>
<td>89.96</td>
</tr>
<tr>
<td>UQI</td>
<td>113.31</td>
</tr>
<tr>
<td>WSNR</td>
<td>135.00</td>
</tr>
<tr>
<td>VIFP</td>
<td>165.53</td>
</tr>
<tr>
<td>MSSIM</td>
<td>175.07</td>
</tr>
<tr>
<td>VSNR</td>
<td>322.46</td>
</tr>
<tr>
<td>NQM</td>
<td>552.10</td>
</tr>
<tr>
<td>IFC</td>
<td>2669.12</td>
</tr>
<tr>
<td>VIF</td>
<td>2716.38</td>
</tr>
<tr>
<td>PSNR Blurring</td>
<td>2.05</td>
</tr>
<tr>
<td>Delta</td>
<td>4.11</td>
</tr>
<tr>
<td>MSAD</td>
<td>4.11</td>
</tr>
<tr>
<td>VQM</td>
<td>28.75</td>
</tr>
<tr>
<td>Blocking Beta</td>
<td>28.75</td>
</tr>
</tbody>
</table>

4.5 Stereoscopic quality metrics

3D quality metrics is a fairly new research area and there are a few recent papers in the literature. Most of these works start with 2D metrics and try to incorporate information about 3D.

In [163], a compound full-reference stereo-video quality metric is proposed which is composed of two components: a monoscopic quality component and stereoscopic quality component. While the former assesses the trivial monoscopic-perceived distortions caused by blur, noise, contrast change etc., the latter assesses the perceived degradation of binocular depth cues only. Structural similarity index is used as a measure for perceptual similarity and a multiscale algorithm is designed for obtaining a perceptual disparity map and a stereo-similarity map to be used in the suggested metric. Block diagram of the metric is given in Figure 4.10. The performance of the metric is verified with subjective tests on distorted stereo images and coded
stereo-video sequences with a final aim to build a perceptually-aware feedback for a H.264 based stereo-video encoder.

**Figure 4.10 Block diagram of the 3DQ metric**

In [164], an extension of 2D-metric which involves the measure of the disparity map distortion was proposed and tested with a methodology for subjective assessment of stereo images. SSIM is enhanced for depth maps using a local or global approach. Block diagram of the local approach for SSIM is given in Figure 4.11.

**Figure 4.11 Block diagram of the proposed 3D metric, adapted from [164]**

In [165], comparison of subjective tests and 2D metrics are given for video plus depth represented 3D video. Results show that the output from the VQM objective metric can be
mapped, so that it correlates strongly with both the overall viewer perception of image quality and depth perception.

In [166], Stereo Band Limited Contrast (SBLC) metric is proposed for evaluating stereoscopic images under compression. It is based on matching the regions of high spatial frequency between the left and right views of the stereo pair and accounting for HVS sensitivity to contrast and luminance changes in regions of high spatial frequency. Matching algorithm uses Sift to extract edges and corners, and the RANSAC algorithm match the regions.

In [167] a metric is proposed to evaluate depth image based rendering (DIBR) for video plus depth video. It is composed of Colour and Sharpness of Edge Distortion (CSED) measure. Colour distortion measures the luminance loss of the rendered image compared with the reference, and sharpness of edge distortion calculates a depth-weighted proportion of remaining edge to the original edge.

In [168], a rapid method which does not use depth map to objectively measure stereo image quality is proposed. Metric has two components: image quality (average of PSNR of two views) and stereo sense (based on the absolute difference between stereo images).

In [169], a metric is proposed for asymmetric coded 3D video using spatial frequency dominance model. Metric is calculated using a weighted sum of difference between images and spatial frequencies’ impurities between degraded and original images.

4.6 Why 2D metrics fail for 3D images

A widely used approach for stereo video quality evaluation has been to apply some 2D metric to evaluate the quality of each video channel separately, and then to declare the overall 3D video quality to the mean of the two figures. While such approach might work for impairments equally affecting the two channels, it would fail in many other cases. Just as taking mean quality over separate video frames would fail to account the effects of temporal masking, any combination of quality estimations over separate video channels would fail to predict the inter-channel masking effects in stereoscopic vision.

The differences between the images of a scene seen by each eye form one of the important cues HVS uses for perceiving depth. The visual system is able to fuse the visual input from both eyes into a single, 3D image. However, artefacts in a 3D video might cause contradictory depth cues to be sent to each eye. There are different theories how HVS deals with such discrepancies, but the common belief is that instead of combining the contradictory cues, image perceived by one eye suppresses the image seen by the other. The binocular suppression theory states that there are masking and facilitation effects between the images perceived by each eye [161]. It is expected that the masking between eyes works similarly to masking effects between different spatial orientations [3]. Experiments with asymmetric video encoding show that when the quality of the two channels differs, the overall perceived quality is closer to the quality of the better channel [162], rather than to the mean quality of the both channels.

Additionally, the image content in both channels also affects the way the quality of each channel would impact the overall 3D quality. For example, the same level of DCT quantization might result in different perceptual quality, based on the depth cues present in a stereo image. In Figure 4.12, both channels of a stereo-pair are compressed with the same quality factor. When an object appears on the same place in both frames, it is equally affected by blocking in each frame, and the perceived cyclopean image is similar to the one shown in Figure 4.12a. When the object has different horizontal position in each frame, the blocking artefacts will affect differently the object in each frame, which results in a cyclopean image similar to the one Figure 4.12b.
The reason for the poor performance of the present 3D quality metrics is that they fail to model the stereoscopic vision in a perceptually correct way. We assume there are two most important HVS properties that should be taken into account when designing 3D quality metric:

I. Separate visual paths for cyclopean image and binocular disparity. HVS is not sensitive to the quality in the left or right channel – instead, it perceives degradation of the cyclopean view as “2D” artefacts, and degradation of the disparity as “3D” artefacts.

II. The overall quality of a 3D scene is a non-linear combination of the “cyclopean” and “binocular” perceptual qualities. Even though “2D” and “3D” artefacts are separately perceived, the content in one visual path might influence the other:
   a. Binocular perception of depth is influenced by pictorial depth cues. It is possible that there is masking or facilitation between depth cues coming from the two visual paths. Consequently, “3D” quality is influenced by “2D” content.
   b. Perception of asymmetric quality depends on the scene depth. Artefacts in the cyclopean view might be masked by the convergence process, as exemplified in Figure 4.12. Consequently, “2D” quality is influenced by “3D” content.

Therefore, a perceptually-correct full-reference 3D quality metric has to implement the following steps: 1) extract cyclopean image and binocular disparity of both test and reference video streams, and compare them separately (assumption I); 2) calculate the unified quality as a functional combination of both measurements, taking into account one visual path influencing the other (assumption IIa and IIb). For comparison of cyclopean images, one can use a quality metric optimised for 2D video. In order to derive the function that combines “2D” and “3D” quality there is a need to have adequate knowledge about the relative importance of various depth cues in order to reflect assumption IIa and there is a need to have adequate knowledge about how varying quality of depth influences the formation of the cyclopean view. In order to gather such knowledge two groups of subjective tests have been designed. The first group investigates the relative importance of depth cues in the particular case of portable autostereoscopic displays. The experimental setting and results are presented in Section 5. The second group of subjective tests investigate the influences of varying depth on the formation of the cyclopean image as well as the influences of varying compression artifacts on the ‘3D’ quality. The design of these tests in presented in Sub-section 6.2.
5 Relative importance of depth cues

5.1 Research method

We have designed and carried out subjective tests to investigate the impact of different depth cues on the perception of depth. The aim of these tests was twofold: first, to estimate the relative importance of binocular depth versus various pictorial depth cues on the depth perception; second, to assess the perceptual influence of video quality degradation over these artefacts and the ability to make depth judgements.

We have created a synthetic 3D scene, where the appearance of different depth cues (i.e. shadows, texture, scaling, focal depth, disparity) can be separately controlled. We rendered multiple images of that scene where different combinations of these depth cues were present. Based on the research in [14], we selected blocking artefacts as the most typical quality degradation in mobile 3D video, and created a test database, containing multiple versions of each image exhibiting different amount of blockiness.

In order to estimate the “3D” quality of each image, we designed task based subjective tests. Instead of asking the test subjects to judge the quality, we asked them to judge the relative depth of an object in the scene, and collected their answers and the time needed for making the decision. In each task, the participant had to select which object appears at the same depth as another object used as reference. Separately, we collected subjective opinion for the visual quality of each presented scene.

Our subjective tests were divided into two parts: performance analysis of the subjective depth estimation, and evaluation of the subjective quality. In the first part, a subjective study was done to collect information on if the participants perceived the depth correctly or not. In the second part, the perceived quality was rated on an unlabelled scale and the acceptance on a binary scale.

ITU-R Recommendations give specification on how to perform subjective video quality evaluation studies [113]. We used Absolute Category Rating (ACR) which is fundamentally a single stimulus method (shown without a reference). According to Pinson and Wolf [96], single stimulus methods can be affected by the context and ratings can drift over time. To compensate for this problem, images were shown several times in random. In the subjective depth estimation test the participants evaluated the relative depth of several objects and gave their answers on a numeric keypad. The chosen method allowed reliable results to be gathered efficiently and analysed easily. A quantitative quality evaluation method was chosen for this work and combined with a binary acceptance rating based on Jumisko-Pyykko et al. [98] in the quality evaluation test. Efficiency and sufficient diversity of the method encouraged the choice.

5.2 Test content

Test images were designed to consist of a constant background scene and a number of objects. As a background of the scene, a room with wooden door and a grey wall was shown. The scene was decided to be simple in order to keep the focus on estimating the depth of simple objects and on evaluating how existence or absence of each cue affects perceived quality of the image. Additional objects on the scene might affect the correct estimation of depth and bias the quality evaluation of images. Evaluating if the estimation was affected by additional objects would be challenging and would require some further studies. Reducing the amount of objects on test images also allows increasing their sizes, which increases clarity of a small image. Objects on the scene were decided to be balls with number labels. Their locations and sizes were randomized in order to avoid hints suggested by the constant locations and sizes of the objects.
Only one of the numbered objects could be on the same depth with the reference object, and there was a fixed range that was kept clear from other objects.

The images were created for portable 3.1" autostereoscopic display with HDDP pixel arrangement created by NEC Technologies [99]. Its input is given side-by-side in resolution of 854x240 pixels, where the left view is given as the left and right view as the right half of the image. Figures 5.1-5.3 illustrate the formatting of side-by-side content.

Six different depth cues were used: focal blur, shadows, texture, binocular view optimal for the 3.1" device the tests were run with, binocular view optimal for a large HD display, and a combination of the first four aforementioned cues. These cues were used in images with three different JPEG compression levels: best quality with no compression, good quality with some slightly visible compression artefacts (90% JPEG compression), and low quality with noticeable blockiness (50% JPEG compression). To illustrate the created content, some images with all used depth cues and compression rates are presented below.

In Figure 5.1 an image with only shadows as a depth cue is presented. The image was not compressed and had no visible blockiness or ringing artefacts. Locations and sizes of the objects were randomized within the image limits. Ball number four is at the same distance as the reference ball marked with letter 'R'. Image was shown in 2D not to give binocular cues to depth, and the two sides of the input to the device were identical.

![Figure 5.1 Image with shadows. Original image with no compression. For reporting purposes the image was compressed (100% JPEG). The ball number four is at the same distance from the viewer as the reference ball marked with 'R'.](image)

In the texture depth cue test, a 2D image with texture as the only cue to depth is introduced. The image is heavily compressed (50% JPEG compression) and the artefacts can easily be seen. Again, locations and sizes of the objects are randomized and the only cue to depth is texture the further the object is, the smaller the texture.

Figure 5.2 represents the last monocular image, having focal depth as the only depth cue. In this type of images, the focal length was defined in such a way that the reference ball is in focus. Thus, the object at the same distance as the reference is the sharpest of the numbered objects (number five), and the other objects are blurred. The further from the reference object a numbered object is, the more blurred it is. In this figure the image is heavily compressed (50% JPEG compression).
Figure 5.2 Image with focal depth and 50% JPEG compression. The ball number five is at the same distance from the viewer as the reference ball marked with 'R'.

Figure 5.3 shows an image with no depth cues other than binocular disparity. The left and right images differ from each other in such a way that when falling on respective retinas, a sensation of depth is created. Ball number four is the one at the same distance as the reference ball. The compression rate is 90%, and some artefacts can be seen.

Figure 5.3 Stereoscopic image designed for a portable screen. 90% JPEG compression. The ball at the same distance from the viewer as the reference ball marked with 'R' is ball number four.

Another stereoscopic example is shown in Figure 5.4. The image is designed for a HD display and scaled down for mobile display resolution. Thus the binocular disparity is smaller and the range of depth in the image should be smaller than in images that are designed straight for mobile use. Again, slight compression is introduced to this image (90% JPEG), and the ball at the same distance as the reference is ball number two. In Figure 5.6 a combination of shadows, texture and focal depth is introduced to a stereoscopic image. There was no compression in the original test image. The ball at the same distance as the reference is number four.
Figure 5.4 Image with all the depth cues, original image with no compression. For reporting purposes the image was compressed (100% JPEG). The ball number four that is at the same distance from the viewer as the reference ball marked with 'R'.

5.3 Subjective tests

5.3.1 Test design

A total of 30 participants were recruited. The recruiting was restricted by age because of the limitations of resources. The participants were with age between 20 and 33, mean age being 25 years. 68% of the participants were males and 32% females, and their attitude towards technology was questioned. They were classified as "early adopters" or "early mainstream", as defined in the Domain Specific Innovativeness Scale [92]. The participants were required to have normal or corrected to normal visual acuity and colour perception, and this was screened before the tests. The depth estimation performance analysis was conducted in order to find out if some depth cues dominate in terms of importance. This was tested with a test where participants estimated depth in similar scenes where only the depth cues and compression levels varied. The other variables were kept as stable as possible during the test. The number of correct answers on the depth estimation task for each pair of independent variables can be seen as an indicator of the importance of this particular pair. As another way of measuring the importance of different independent variable pairs, speed of carrying out each depth estimation task was recorded. This was expected to give a straightforward hint of the easiness of conducting each task, and thus stating which combinations made the depth easy to estimate. Even if the answer was incorrect, the task might have seemed to be easy with some depth cues, and the speed was recorded to find such cases. In the quality evaluation, the aim was to find out if the scene with a certain depth cue was also pleasant to watch. Keeping other parameters constant, the presence or absence of depth cues can give hints about their effect on pleasantness. As in the depth estimation part, the effect of reducing the quality was also studied.

Satisfaction on a binary scale shows more precisely what sort of deficiencies on the images is still acceptable. The acceptability rate together with the satisfaction score gives a threshold of acceptance. This part of the study is expected to answer the following questions:

- Are there some depth cues that make the scene more pleasant and acceptable?
- Are some depth cues more vulnerable to suffer from compression?

5.3.2 Test procedure

Test room illumination was fixed before starting the test. Test conditions for viewing have been suggested by recommendations of ITU-R BT-500 [113]. The suggestions were given for a 17" screen, and since no standards have been given for the viewing conditions for a portable display, the suggestions were adapted to suit mobile viewing. The adapted mobile viewing conditions in terms of illumination and viewing distance are illustrated in Figure 5.5.
The device that was used in this study is a prototype of an autostereoscopic display created by NEC LCD Technologies. Size of the display is 3.1” with resolution of 427x240 pixels. It is backwards compatible with 2D content and has the same resolution in 2D and 3D modes. The switch between the modes can be done with software means and different parts of the scene can even be in different modes simultaneously. Earlier tests with the device have shown that it offers excellent stereo effects with great angle of view and low crosstalk. Participants were asked to take a good and relaxed position on the seat, and after this the viewing distance and angle were adjusted according to the participants' personal preferences. The viewing distance was 40cm. In the first part of the experiment the aim was to get information about the efficiency and effectiveness of the evaluation of depth.

A quality evaluation study was conducted, aiming to gather information about the participants' satisfaction at each image. By taking the aforementioned factors into account, the study follows the guidelines given by the ISO 9241-11 standard [100], which states that the most important parameters of usability are effectiveness, efficiency and satisfaction at a defined context of use.

Pilot tests were conducted before the experiments in order to fix the length of the test sequences and good compression rates. In the pilot tests it was found that the amount of images in the depth estimation task should not be as big as was originally planned. The amount was decreased to 72 in order to keep the duration of the test within 20 minutes. Before starting the test, the participants were given written and spoken instructions in their preferred language, German or English. All the participants could not be chosen within the same language group because of the restrictions set by the unavailability of native German translators and interpreters.

The demographic backgrounds of the participants were collected with a questionnaire and structured questions regarding their occupation, education, attitude towards technology, and experience on 3D video and mobile television. The participants' visual acuity was screened. For screening of vision, hyperopia was tested with a Snellen Chart. Colour vision was tested using Ishihara test, which is a test designed for colour deficiencies [101]. Stereo vision was tested using Randot Stereo Test. The test involves figures on depths from 400 to 20 arcsec. Myopia was not tested, because for a mobile display it is not considered relevant. To find out if the 3D images caused motion sickness side effects, a simulator sickness survey was done using a Simulator Sickness Questionnaire (SSQ) [102], [103]. The participants were given a short oral
introduction to the test procedure, and shown a training set. The training set consisted of 12 images, chosen from the test material so that it represented it extensively. The goal of the training was to get familiarized with the test material, viewing images on autostereoscopic display, and to find the optimal viewing distance.

In the depth estimation test, the participants were asked to evaluate which of the numbered objects in the test images were at the same distance with the reference object. The answer was given by pressing a corresponding number on a numeric keypad. The given answer and the answer time were recorded. A quality evaluation study was conducted, aiming to gather information about the participants' satisfaction in each image by asking their assessment about the quality acceptance and overall quality according to their opinions on an unlabelled scale from 0 to 10. This kind of a scale was chosen to go around the end-avoidance problem [104]. The participants were also asked to state whether they would accept the quality or not. The acceptance of quality was given on a binary scale by selecting "Yes" or "No". Method of acceptance ratings was adopted from Jumisko-Pyykko et al. [98].

Seeing 3D content on a test display and then switching to a normal 2D display to give the answer and working in a normal 3D world in between these two tasks could be confusing or even bias the results because of the comparison of 2D display versus 3D. According to Reiter et al. [97], simple parallax tasks such as pressing a button does not impact on quality requirements. It can be assumed that giving answers on an answer sheet during the quality rating should not have an impact on quality ratings. Thus, pen and paper were chosen to be the method of collecting the quality ratings. For depth estimation task, a separate numeric keypad was chosen for collecting answers.

In the depth estimation part, the participants were shown all the combinations of independent variables four times. This means that the total number of different test images was 36 (6 cues x 3 compression levels x 2 sets). All of them were shown twice, making the total amount 72. This amount of images was agreed after the pilot tests implying that a third image set (in total 108 images) of images would increase the total duration of the study too much. For the quality evaluation, the same images were shown again twice (in total 72 images). Images were shown in randomized order.

5.3.3 Simulator Sickness Questionnaire

In tests, which involve 3D images or virtual reality imagery, motion sickness can occur [103]. A study by Shibata et al. [105] indicates that small screen size might have an impact on feelings of discomfort when viewing 3D images.

In our experiments, we used as a relative measurement tool to eliminate the impact of possible symptoms the participants might have had prior to the test. The SSQ was filled out by the participants before the experiment and straight after the test. To study how the possible symptoms change in function of time, the test was repeated every four minutes until 12 minutes had passed since ending the experiment.

When 12 minutes had passed, the participants' physical state had already returned to the initial level. Right after the test, increased disorientation and oculomotor related symptoms were reported. There was one outlier who reported significant effects on increased disorientation and oculomotor related symptoms, but also nausea. After eight minutes, it was hard to say which caused the symptoms, as the feeling of nausea had slightly increased, but the oculomotor related symptoms had decreased to a level that was lower than prior to the test.
5.3.4 Test automation

A script to show the images and collect the answers was created with MATLAB. The program is divided into three parts: Training, Depth estimation and Quality evaluation, as shown in Figure 5.6.

![Figure 5.6 A block diagram of the test procedure](image)

In the Training part, the program shows a pre-defined set of training images in a randomized order. A button press from the keyboard starts the sequence. The next image is displayed always when a button is pressed, until all the images have been shown. Pressing the Escape button escapes the sequence and starts running the program again. The sequence ends with a black screen. Pressing the Escape button on the black screen escapes the whole program.

In the Depth estimation part, there are two different pre-defined, randomized play lists, from which one is shown. From the black screen, a button press starts the next sequence. At the same time when an image is displayed, a timer is started. Pressing a numbered button from one to five records the answer and stops the timer. Only numbers from one to five are approved, otherwise the timer keeps on running and nothing happens. Again, the sequence can be escaped and started from the beginning by pressing the Escape button. The sequence is shown twice.

The last part is Quality evaluation part. The same images are shown as in the previous part, and the impulse of showing the next image is given by a button press. Any button is allowed, except for the Escape that again escapes the sequence and starts it again from the beginning. Only the shown play list is recorded.

5.4 Results

5.4.1 Speed of Depth Estimation

Before the analysis, the speed results were normalized per user in order to have data that is valuable for comparisons. For each user, all the used times were scaled between zero and ten.
The data analysis follows the methodology proposed by Coolican [106]. Friedman's test was chosen to measure differences between several data sets [106]. Wilcoxon test was selected for analyzing the differences between the compression levels. In order to compare the results of depth estimation speed with different depth cues and to compare their relative significance. A total of 15 comparisons were made by each participant. In order to lower the risk of making the Type I Error, Bonferroni correction was used and a new significance level was \( p < 0.0033 \). Bonferroni correction was used also when comparing the speed on different compression rates. The used significance level was \( p < 0.0167 \).

A Friedman test indicates a significant difference between the times used on depth estimation task on three different compression levels (\( X^2 = 10.117, \ p < 0.05 \)). A Wilcoxon test indicates a faster conduction of depth estimation when the image is not compressed at all compared with the lower quality images. The difference between the best and good quality is significant (\( Z = -2.644, \ p < 0.0167 \)), and between the best and the low quality images even more remarkable (\( Z = -3.815, \ p < 0.0167 \)). There is no significant difference between the results for the compressed images, and the results shown in Figure 5.7 indicate that any compression affects the speed of estimating depth.

![Figure 5.7](image)

**Figure 5.7** The means of the times used for the depth estimation task. The time values have been normalized per user to only show the differences. The bars represent the Best, Good and Low perceived quality.

A Friedman's test for different depth cues indicates that there is a significant difference between the variations of the data sets (\( X^2 = 140.871, \ p < 0.05 \)). A pair-wise Wilcoxon test was done to study further how the data sets differ from each other. The results indicate that for the images with stereoscopic depth cues the recorded times are faster than for the images with only monocular cues. Between the stereoscopic cues there were no significant differences, and the same applies for the monocular cues, as seen in Figure 5.8. The significance level is \( p < 0.0033 \).
5.4.2 Correctness of depth estimation

For analysing the results for correctness of depth estimation, McNemar’s test with Bonferroni correction was used. The p-values were $p<0.0167$ for variable compression rates, and $p<0.0033$ for variable depth cues.

The results from the McNemar test indicate that there is a difference between the correctness of ratings in the low and the good quality images ($X^2=23.223$, $p<0.0167$), as well as between the low and the best quality images ($X^2=23.770$, $p<0.0167$). No significant difference on the correctness of ratings was found between the good and the best quality images, as seen in Figure 5.9. This might imply that depth can be perceived correctly even if the quality is degraded by moderate compression. However, for lower quality images the difference is quite pronounced.

When testing the differences between the correctness of ratings on different depth cues, a McNemar test indicates that the images with stereoscopic cues gave more accurate results than the images with only monoscopic depth cues. No significant difference could be found between different stereoscopic depth cues, and the same applies for monocular depth cues, as seen in

---

Figure 5.8 The means of the times used for images with different depth cues

![Figure 5.8](image)

Figure 5.9 Correctness of results of the depth estimation task on different compression levels. The bars represent the Best quality (no compression), Good quality (90% JPEG compression) and Low quality (50% JPEG compression) from left to right.

![Figure 5.9](image)
Figure 5.10. The significance level in use was $p<0.0033$. Spearman’s test indicates that the times used for estimating the depth and the correctness of the result have a negative correlation ($r=-0.148$, $p<0.01$).

![Bar chart showing correctness of answers for different depth cues](image)

Figure 5.10  The correctness of results of the depth estimation task for images with different depth cues

5.4.3 Quality Ratings

Non-parametric methods were selected to analyze the quality ratings data, since the distribution of the data did not fulfill the requirements of normality ($Kolmogorov-Smirnov > 0.05$). As in the previous part of the analysis, Friedman’s test was used to analyze the variance of several data sets, and Wilcoxon test was used to find differences of two data sets pair-wise [106]. The same Bonferroni corrected significance levels were used ($p<0.0033$ and $p<0.0167$). A Friedman’s test indicates a significant effect for the qualities given on different levels of compression ($X^2= 403.5$, $p<0.05$). A Wilcoxon signed rank test indicates a significant preference of the images with lower compression. The Best quality was preferred over Low quality ($Z=-16.456$, $p<0.0167$), and over Good quality ($Z=-10.082$, $p<0.0167$). Good quality was preferred over Low quality ($Z=-12.287$, $p<0.0167$). A Friedman's test for the qualities given for images with different depth cues indicates that some data sets differ from each other significantly ($X^2 = 420.944$, $p<0.05$). A Wilcoxon signed rank test indicates that the images with depth cue Dnsscaled were rated with highest scores, and followed by Mobile and Shadows. The two latter ones had no significant difference between them. Then, the images with textured objects were preferred over the ones with focal depth and the combination of all cues, as seen in Figure 5.11. The significance value used was Bonferroni corrected, $p<0.0033$. 
5.4.4 Acceptance Ratings

A Spearman's rank test indicates a correlation between the qualities and the acceptance values ($r=.707$, $p<0.01$). To find the threshold for acceptance ratings, a method by Jumisko-Pyykö et al. [98] was used. The threshold was searched by connecting the binary satisfaction data with the quality ratings. The threshold was determined based on the statistics of the acceptable (Mean: 7.66, SD: 1.77), and the unacceptable quality ratings (Mean: 4.0, SD: 1.68). The threshold value was defined to be approximately 5.8.

A Chi-Square test of independence showed a significant effect on acceptance data between the data sets. A McNemar test between the cases showed that for all the three different compression levels, the effect was significant. The low quality images are in the category of unacceptable (Mean: 5.03, SD: 2.079). The acceptance ratings for different compression ratios are given in Figure 5.12.

![Figure 5.11 Quality ratings for images with different depth cues.](image-url)
For the acceptance of different depth cues, the McNemar test indicates that the downscaled images got the best acceptance ratings. The ratings given for the images with shadows and images designed for a portable display are only little lower than the ones given for downscaled images, and cannot be distinguished from each other. The images with texture were only barely acceptable. The images with "All Cues" and "Focal Depth" as depth cues are regarded as unacceptable. Based on the current experiment data, it cannot be determined whether the combination of all cues got low ratings because of the strong influence of the focal depth. The acceptance ratings for different compression ratios are given in Figure 5.13.

Figure 5.12 Acceptance ratings for three different compression ratios: Low quality (50% JPEG compression), Good quality (90% JPEG compression), and the Best quality images (no compression).

Figure 5.13 Acceptance ratings for the six different depth cues. The used cues from left to right are as follows: a combination of all cues, downscaled stereoscopic image, an image with focal depth, a stereoscopic image designed for a mobile display, shadows, and texture.
6 Towards stereoscopic HVS modelling for feature-based 3D video quality evaluation

6.1 Relative impact of 2D and 3D depth cues on overall 3D perception

According to previous studies, motion parallax and stereopsis are considered the most important depth cues [108]. Our results on relative importance of depth cues, as presented in Section 5, support this theory. All images with stereoscopic depth cues got significantly better ratings than the images with monocular cues when measuring the accuracy and speed of depth perception. The finding that additional cues to depth (shadows, texture, and focal depth) did not lead to better results might be an implication of the supremacy of stereopsis over monocular depth cues on portable displays. In particular, focal depth seems to get a low ranking in all the tested areas: in accuracy, efficiency and in acceptance. In informal discussions after the experiments, several participants complained that the images with focal blur appeared out of focus. Our conclusion is that depth of field does not work well as a separate depth cue on a mobile display observed from close distance, and should be used with extra caution. Blurring is perceived mostly as a “2D” artefact, and affects the cyclopean image of a 3D scene, rather than its binocular disparity.

The compression level had a significant effect on both the accuracy and the speed of depth perception. The results of our subjective experiments indicate that the quality of the image has a significant effect on the correctness of depth estimation, but an even greater effect on the speed of estimation. The quality might have some imperfections and the depth might be still perceived perfectly, but as the quality drops further, so does the accuracy of depth perception. The time used for estimating the depth and the correctness of the estimation had a negative correlation, indicating that if the depth estimation was conducted fast, the result was also most likely correct.

Interestingly, the effectiveness and accuracy of the depth estimation was not consistent with the results of the quality evaluation studies. A previous study by Strohmeier et al [109] indicates that the depth effect only contributes to the quality perception when there is little compression. This kind of effect can be easily understood, especially in case of 3D video: if lower quality affects the efficiency of depth perception, the binocular HVS might not have enough time to create a plausible 3D effect for moving objects. On low bit rates, the benefits gained by added depth might be invalidated by the artefacts. This gives further support to the assumption that 3D perception is an extremely delicate mechanism and vulnerable to artefacts. This effect suggests there is an influence of 2D artefacts over the 3D perception. A scene with a good “3D” quality, (i.e. which allows fast depth judgement), might be graded as having low overall quality due to 2D artefacts.

Another interesting outcome is that scenes with the same amount of blockiness (same compression rate for each channel) but with different depth levels (i.e. 2D, “downscaled” 3D and “optimal” 3D) were graded as having different quality. This effect implies that different amount of 3D depth cues (binocular disparity) influences the perception quality.

The results of the study, described in Section 5 form a solid basis for further developments. However, more subjective tests are needed to quantify the mechanisms of mixing ‘2D’ and ‘3D’ qualities. The design of a subsequent such study is described in Sub-section 6.2.

6.2 Impact of binocular disparity on overall quality

The perceived quality of a 3D scene is a combination of two components – “2D” quality (image details), and its “3D” quality (quality of the binocular depth cues). However, the very presence of stereoscopic depth changes the way image details are perceived. Equivalent loss of image fidelity is receives different quality ratings in 2D and 3D video [109]. There are two ways
binocular disparity affects the “cyclopean” quality, depending on whether the disparities are considered local or global.

Local binocular disparity changes eye convergence, which determines which image regions are fused together in the cyclopean view. Stereo-pair with asymmetric quality will be perceived in a different way based on the position where artefacts appear in left and right channel. The principal problem is to find a plausible way image patches are fused together to form a relevant cyclopean view. Furthermore, the model of binocular rivalry should be tuned to match subjective quality scores for asymmetric stereo views.

Global binocular disparity affects the perception of volume in a 3D scene. In general, larger variance of disparities in a stereoscopic pair, generate a more “spacious” perceived scene. The subjective experiments in [109] show that the presence of depth influences the perceived quality, and the influence can be either positive or negative depending on the content. The subjective experiments described in Chapter 5 demonstrate that the same amount of blockiness is graded differently in scenes with differently pronounced depth. The influence is not monotonic, which implies there might be an “optimal” global depth of a 3D scene on portable auto-stereoscopic displays, for which HVS has lowest sensitivity to cyclopean image degradations.

In order to model the quality of a 3D scene, a quality metric should be able to predict the impact of local and global binocular depth on the overall perceptual quality in the presence of varying compression artefacts. A denser scale of varying depths and a denser scale of varying compression artefacts are needed.

We have designed a set of subjective experiments, where the binocular depth and the image quality of a scene are independently varied on denser grids. In contrast to the previous study, where synthetic contents was used, in the current tests, four real-world multiview videos were used instead. From each multiview sequence, a number of stereoscopic videos were created using different camera pairs, thus achieving varying camera baseline for the same scene. More specifically, the depth was varied between 2D, HD-optical baseline and mobile-optimal baseline. Each stereoscopic video was compressed with varying quality, i.e. with five different QPs. In the planned tests, the participants will be asked to grade the quality of each compressed video in a single stimulus setting. Thus, we expect to get enough data from subjective tests to design and tune the blocks related with the formation of the cyclopean view and with the disparity estimation and their combination.

6.3 Image processing channel for stereoscopic quality estimation

In order to follow closely the subjective quality scores, an estimation of 3D quality should take into account the specifics of the stereoscopic depth perception. Even if feature based, a 3D quality metric should be “3D model-aware”, and follow the stages of stereoscopic vision.

As we discussed in Chapter 2, the visual information collected by each eye, first passes through lateral geniculate nucleus (LGN) before reaching the V1 visual centre in the brain cortex. The stages of the stereoscopic vision can be modelled by an image processing channel. Figure 6.1 shows a side-by-side comparison of a simplified block-diagram of the 3D vision (top) and the processing channel modelling it (bottom).
The visual information is collected by the photoreceptors in the retina or each eye separately. The retinal processing is responsible for a number of specific properties of HVS, which can be modelled as luminance masking, colour processing and contrast sensitivity, as discussed in Chapter 3. The luminance, colour and contrast adaptation occur in each eye separately. After that, both optical nerves arrive in the LGN. LGN decorrelates the visual information, greatly reducing the visual information – the number of outgoing visual nerves is only 1% of the number of neurons going to the LGN. The processes of binocular masking and extraction of binocular depth cues happen are performed in LGN. After that, a single, fused representation of the scene (also known as “cyclopean image”) is fed to the V1 brain centre. The processes in V1 are modelled as multi-channel decomposition, masking between channels with different spatial frequency and orientation, and finally, temporal sensitivity and masking. The binocular suppression theory and also anatomical evidence suggest small part of the visual information
delivered from each eye might be fed directly to V1 without being processed by LGN. The depth perception occurs further, in visual processing centres after V1, and it is a combination of cues extracted by LGN, separate visual information from each eye, and higher level visual cognition, involving perspective, shadow, and other depth cues.

We believe that this order of visual processing makes possible also masking and facilitation effects between different depth cues. The presence and strength of one type of depth cues might suppress or enhance the perceptibility of another. This would necessitate that depth cues and binocular masking are estimated together, in the same processing block.

As discussed in Chapter 4, feature-based quality estimation is more suitable for specific applications, with limited set of expected artefacts and computational complexity constraints – which is the case for Mobile 3D broadcasting over DVB-H. Aiming at feature-based quality metric, which is also sufficiently ‘model-aware’, we propose the image processing channel shown in Figure 6.2. The channel operates in a Full Reference (FR) mode, comparing the visual difference between a test and reference 3D video streams.

First, each channel is separately transformed into a perceptual colour space. Good candidates for such colour space are S-CIELAB [65], which models colour processing, luminance and contrast masking, and ST-CIELAB [53], which additionally models temporal masking. The next block estimates the visibility of depth cues, and eventual masking between different depth cues. An optional block would estimate the temporal integration and masking – and models, used in [54] and [56] are suitable to be used in this block. Finally, the processed visual information of the test and reference video streams is compared by a “structural similarity” metric, which estimates CSF and perceptual salience of different areas in the visual content. Good algorithms for structural similarity comparison are presented in [142] (based on local statistics), and in [146] (using adaptive scales).
A cyclopean view and a perceptual disparity map are derived separately for the test and the reference video stream. Suitable algorithms for this block are to be designed. The perceptual similarity between test and reference cyclopean view is used as an indicator for “2D” quality impairment. Statistical comparison of test and reference disparity maps (for example, normalized difference or average absolute deviation) is used as indicator of “3D” quality degradation.

Finally, a non-linear weighting function should be used to combine the “2D” and “3D” quality measurements into one compound metric $Q_{\text{total}}$:

$$Q_{\text{total}} = Q_{\text{2D}} + Q_{\text{3D}} \text{ where } k = f_{\text{depth}} Q_{\text{3D}}, \ l = f_{\text{cues}} Q_{\text{2D}}$$

where $f_{\text{depth}}$ is a function which describes the influence of binocular disparity over the “2D” quality, and $f_{\text{cues}}$ is a function which describes the influence of cyclopean view artefacts over the depth perception. For approximating $f_{\text{depth}}$ and $f_{\text{cues}}$ weighting functions the outcomes of the subjective experiments, described in Section 6.2 will be essentially used.
7 Conclusions

This technical report addressed the topic of modelling the stereoscopic HVS for the needs of a future feature based objective 3D quality metric. The physiology of the vision has been overviewed along with models of HVS and state-of-the-art visual quality metrics. A brief analysis about why 2D metrics fail in giving adequate results for 3D images and video has been presented. It helps in arguing the need of two specific subjective studies run in a sequence. In the first study, the relative importance of 3D visual cues on a portable auto-stereoscopic display has been addressed. The study has emphasized the role of the binocular cue for perceiving the depth and the influence of the varying depth on the overall quality. The second study, being in progress, has addressed the latter problem more thoroughly by aiming to quantify the influence of the, possibly distorted, depth on the formation of the cyclopean view and the role of 2D compression artefacts on both the cyclopean view and the stereoscopic ‘delta’. In both studies, the specifics of mobile 3D display have been taken into account. The main outcome of the report is an image processing channel for feature-based 3D quality evaluation, incorporating the essential knowledge about the stereoscopic HVS. While rather general, it forms a solid basis for developing an objective 3D video quality metric in line with the findings from subjective tests.

References

[18] Howard, B. Rogers, “Binocular vision and stereopsis, 1995
[27] M. Bradshaw, B. Rogers, “The interaction of binocular disparity and motion parallax in the computation of depth”, Vision Research, 1996
[41] E. Peli, “In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths”. Vision Research, 37 (23), 1997
[56] [S. Winkler, “A perceptual distortion metric for digital color video”, in Proc. of SPIE 3644:175-184, San Jose, CA, Jan 1999.


[119] "A57" database, available online at: http://foulard.ece.comell.edu/dmc27/vsnr/a57_db.zip


MOBILE3DTV - Mobile 3DTV Content Delivery Optimization over DVB-H System - is a three-year project which started in January 2008. The project is partly funded by the European Union 7th RTD Framework Programme in the context of the Information & Communication Technology (ICT) Cooperation Theme.

The main objective of MOBILE3DTV is to demonstrate the viability of the new technology of mobile 3DTV. The project develops a technology demonstration system for the creation and coding of 3D video content, its delivery over DVB-H and display on a mobile device, equipped with an auto-stereoscopic display.

The MOBILE3DTV consortium is formed by three universities, a public research institute and two SMEs from Finland, Germany, Turkey, and Bulgaria. Partners span diverse yet complementary expertise in the areas of 3D content creation and coding, error resilient transmission, user studies, visual quality enhancement and project management.

For further information about the project, please visit www.mobile3dtv.eu.