# Assessing human depth perception for 2D and 3D stereoscopic images and video and its relation with the overall 3D QoE

vorgelegt von
Master of Science
Pierre Lebreton
aus Le Mans

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
Assessment of IP-Based Application
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Sebastian Möller
Erstgutachter: Prof. Dr. -Ing Alexander Raake
Zweitgutachter: Prof. Dr. Ingrid E.J. Heynderickx
Drittgutachter: Dr. -Ing Marcus Barkowsky

Tag der wissenschaftlichen Aussprache: 11.12.2015

Berlin 2016

D 83

# Acknowledgements

My journey to the PhD would not have been possible without the guidance of great people who showed me the way of research. There have been some major “push” which drove me to science and then to the PhD. First, I would like to thank Patrick Le Callet to which I owe my career as a researcher. He was the first one who gave me the opportunity to do research, guided me during my first experiences in research, supported my applications for my masters thesis at NTT, and then my PhD at the technical University of Berlin. I would not have been where I am now without your kindness and great support.

I would also like to thank Alexander Raake, first for agreeing to supervise my thesis, but mainly for all the fruitful discussion, his kindness, continuous support along the thesis. Each meeting gave me new pikes of energy enabling me to go further, see new aspects that I may have missed, and new ideas. This was always done constructively with respect of my research interest. It really has been a great pleasure to work with you.

Also a very important contribution to my thesis, but also to my career as a scientist is Marcus Barkowsky. I would like to thank him for the time he took during my entire thesis to advise me about my research, my experiments, the analysis of data. I really learned a lot, learned how to be critical of my work, see the limits of our results and analysis. Even if sometimes it was more difficult, it was really a process I needed to learn to do better work. There is still much to learn, but I know the direction.

I also would like to thank Akira Takahishi and Kazuhisa Yamagishi who also have their stone in the construction of my thesis. Not directly, but were very kind to accept me at their laboratory, guided me though my first subjective experiments, and along my masters thesis. This gave me the willingness to pursue to the PhD.

And of course, I would like to thank my father for supporting me in doing a PhD. Even though Germany is not that far from France, I have been away for a very long time. So thanks again for supporting me. Thanks also to my sisters, my family, and my friends for always being there (and thanks again the ones who made all the way to Berlin !).

# Contents

# Acronyms

| | |
|---|---|
| ANOVA | Analysis of variance |
| ACR | Absolute category rating |
| BT | Bradley-Terry |
| DCT | Discrete cosine transform |
| DERS | Depth estimation reference software |
| DOF | Depth of field |
| EC | Evaluation concept |
| FLMP | Fuzzy logical model of perception |
| GLP | Global layout properties |
| GOP | Group of picture |
| HDMI | High-definition multimedia interface |
| HRC | Hypothetical reference circuit |
| IQ | Image quality |
| ITU | International Telecommunication Union |
| JND | Just noticeable difference |
| LCD | Liquid cristal display |
| LDV | Layered depth video |
| LSD | Line segment detector |
| MAP | Maximum a posteriori |
| MANOVA | Multiple analysis of variance |
| MVC | Multi view coding |
| MVD | Multi view plus depth |
| MLE | Maximum-Likelihood Estimation |
| MWF | Modified weak fusion |
| NANOVA | N-Way analysis of variance |
| OR | Outlier ratio |
| OT | Object thickness |
| PC | Paired comparison |
| PCA | Principal component analysis |
| PVS | Processed Video Sequence |
| QoE | Quality of Experience |
| QP | Quantization parameter |
| RANSAC | Random sample consensus |
| RMSE | Root mean square error |
| RODR | Region of depth relevance |
| RTP | Real-time transport protocol |
| SAMVIQ | Subjective assessment method for video quality BT.1788 |

| | |
|---|---|
| SI | Spatial information |
| SRC | Source reference circuit (ANSI/ATIS adopted by VQEG) |
| TS | Transport stream |
| TI | Temporal information |
| 3DAV | 3D added value |

# Chapter 1
# Introduction

3D was planned as the next step for television. However, it apparently did not manage to convince a large number of end users to equip themselves at their home. In the case of movie theaters, 3D is still receiving attention from spectators and movie producers. One of the important issues to improve the acceptance of 3DTV is to demonstrate the added value of 3D to the user. The contribution of 3D was claimed by the industry to be at the same level as the transition from monochrome to color.

Along this thesis it will be described how 3D can improve the user experience compared to 2D videos, and how the added value of 3D, the perceived depth information, can be characterized. Different aspects about how 3D videos are perceived and how different factors such as the texture quality and the perceived depth affect the Quality of Experience (QoE) will be analyzed.

The following chapter 2 begins with the state of the art on the definition of QoE. It addresses previous results on the interaction between the different factors affecting QoE: image quality, depth quality and quantity, visual discomfort, and QoE. However, since evaluating 3D QoE itself is not straightforward, this chapter will address how it can be possible to measure QoE using different evaluation concepts. The relationship between low-level factors such as image quality, depth, and visual comfort will be put into relation with these high-level evaluation concepts.

Since the perceived depth may represent the added value compared to 2D videos, one major goal of the chapter will be to provide an in-depth discussion of depth perception. Different depth cues will be addressed, analyzing how these depth cues relate to each other. As will be discussed in more detail, the different depth cues are not necessarily orthogonal. Hence the possible interaction between different depth cues will be discussed as well. In addition, different models of depth cue pooling will be addressed, which target the prediction of an overall depth judgment.

Finally, considering that the aspects of perception addressed in most of the reported studies reflect human visual perception in a natural environment, technology factors should also be considered analyzing their impact on what is seen by the participants.

The third chapter addresses the work which has been performed in this thesis by means of subjective experiments, to investigate how it is possible to reveal the added value of 3D over 2D. Different 3D video streaming scenarios are addressed. These scenarios involve error-free or non-error-prone transmission chains. It will be illustrated that in the particular case of the evaluation of 3D videos encoded at different bitrates, it is not easy to evaluate the added value of 3D compared to 2D. Based on this observation, the issues of the understanding of the rating scales by the participants is discussed. Moreover, considering the similarities between the 2D and 3D ratings, the chapter also describes studies on evaluating the performance of 2D video quality prediction algorithms for 3D video quality prediction. Finally, it is shown that by means of paired comparison it is possible to measure the added value of 3D. The preference of 3D over 2D is found to be content-dependent, but also linearly dependent on the image quality resulting from codding. Considering the performance of 2D video quality prediction algorithms for predicting 3D image quality, a key remaining aspect for measuring the added value of 3D compared to 2D is the depth-specific *characterization* of 3D video sequences.

Chapter 4 describes the work performed by the author on characterizing the depth in 3D video sequences by means of subjective testing. As described in section two, there are two different kinds of depth cues: monocular and binocular ones. Both kinds of depth cues will be addressed in this chapter. However, this chapter shows that evaluating depth cues may not be an easy task for the participants. Therefore work has been carried out on how to properly define the scales on which the participants rate the depth in images and videos. Particular attention has been paid to the selection of natural images in order to cover different amounts of monocular and binocular depth cues in the tests. Moreover, in addition to providing a definition for the scales and selected images, research was focused on defining subjective assessment methods for the evaluation of depth cues. The presented result shows that the proposed methodology enables test participants to better understand the task by means of a more intensive training phase which allows test participants to have more examples of what they should do during the test. Finally, the relationship between monocular depth cues and overall depth is studied in the case of natural images.

Based on the subjective ratings collected from test participants using the approach developed in this thesis, Chapter 5 describes the work performed on developing prediction algorithms for evaluating depth cues in natural images. Different algorithms are described to evaluate binocular and monocular depth cues. These algorithms address: binocular depth cues, linear perspective, texture gradient, defocus blur, and in case of video, motion parallax. Similar to the work carried out on subjective assessment methods, the question of the reliability of the prediction provided by these algorithms is addressed. By means of temporal consistency analysis, image classification, and feature analysis on the different algorithms, it is described how it is possible to estimate the confidence of the metrics.

Finally, Chapter 6 reviews the main contributions of this thesis and addresses the perspectives for a continuation of this work.

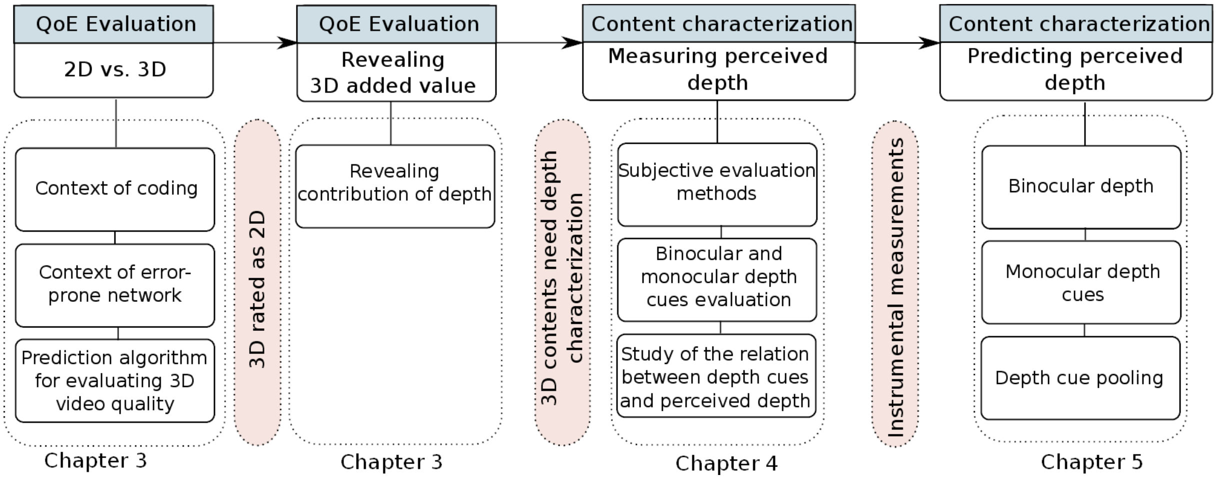

QoE Evaluation
2D vs. 3D
QoE Evaluation
Revealing 3D added value
Content characterization
Measuring perceived depth
Content characterization
Predicting perceived depth
Context of coding
Context of error-prone network
Prediction algorithm for evaluating 3D video quality
3D rated as 2D
Revealing contribution of depth
3D contents need depth characterization
Subjective evaluation methods
Binocular and monocular depth cues evaluation
Study of the relation between depth cues and perceived depth
Instrumental measurements
Binocular depth
Monocular depth cues
Depth cue pooling
Chapter 3
Chapter 3
Chapter 4
Chapter 5

Figure 1.1: Different items studied

# Chapter 2
# State of the art

## 2.1 Introduction

The purpose of this chapter is to introduce the reader to the different relevant aspects of 3D Quality of Experience (QoE). It provides information on the existing results on the different factors involved in the notion of 3D Quality of Experience.

A widely accepted definition of Quality of Experience is provided in the Qualinet white paper [1] :

**Listing 2.1: Definition of Quality of Experience**

```
Quality of Experience (QoE) is the degree of delight or annoyance of the user of an
    application or service. It results from the fulfillment of his or her expectations
    with respect to the utility and / or enjoyment of the application or service in the
    light of the user’s personality and current state.
```

Different aspects are important in that definition: QoE relates to the user’s perception, his expectations, and also depends on a particular context. As stated by Seuntiëns [2], in the context of 3D video presentation different factors drive the overall QoE, e.g. the picture quality, the visual discomfort and the perceived depth. All these three factors are assessed based on the user’s expectation and context of use.

### *2.1.1 Evaluating Quality of Experience*

Evaluating the overall QoE is challenging. First, different factors affect how 3D image and video material is perceived. These factors include the pictorial quality, which relates to the quality of the two stereoscopic pictures seen by the user at a specific instant. When compared to 2D video, other factor has been added with 3D video, namely: the perceived depth, which is the added value of the 3D images and videos. It will depend on both how the image was created, but also how it is viewed: the size of the display used to render the 3D images/videos, the viewing distance, the display resolution, etc. And finally, the third main factor is the visual comfort. Depending on how the content was created and rendered, observers can feel different degrees of stress, and in the long-term suffer from fatigue when watching 3D material. This will affect their overall assessment of the 3D movie/image viewing experience. All these three different factors together combine to the overall 3D experience. Studies have been conducted to evaluate this overall experience and how 2D and 3D quality of experience differ. Results show only little and non statistically significant differences between quality ratings for 2D and 3D uncompressed video sequences in the context of subjective tests with different coding conditions with hidden reference [3, 4]. However, these small differences in quality ratings do not mean that the user’s experience is not different between the 2D and 3D presentation, but rather that the reported quality values may be too much influenced by the context of the subjective experiment, where many different degradations of pictorial quality are presented and this drives the overall quality ratings. This problem is aggravated by the single stimulus rating paradigm. Hence, users provide ratings of pictorial quality rather than of QoE. To tackle this issue, Seuntiëns [2], considered new evaluation concepts such as:

- *Presence* related to the feeling of "being there and reacting to" as defined by IJsselsteijn [5].
- *Naturalness* related to "what observers perceive as a truthful representation of reality" (Ijsselsteijn, [6]).
- *Viewing experience* as defined by Seuntiëns is considering a higher degree of imagination similar to *Presence*. However, it considers that the persons "know they are not in the movie but react in a physical and emotional sense to the story".

These evaluation concepts were put into relation to lower-level factors such as the quality of the representation, the visual discomfort and the depth (Figure 2.1), and it was shown that the viewing experience and naturalness were rated relatively similarly to image quality, and presence ratings were more closely related to depth scores. In addition it was also observed that the type of content, still images or video sequences, can have an effect on the subjective scores. *Viewing experience* and *naturalness* were not affected by the type of content, but *presence* was. Additionally it was observed by IJsselsteijn et al [7] that motion had a much higher impact on *presence* scores than depth, which makes *presence* a less appropriate evaluation concept for 3D videos. Further studies from Seuntiëns comparing *naturalness* and *viewing experience* showed that *naturalness* is the most sensitive metric for measuring "the 3D added value".

Once an appropriate evaluation concept identified, it is possible to investigate the effect of the different "low level factors" (level 2 in Figure 2.1) to the overall 3D QoE. In the following subsections, the relationship between these factors and QoE will be discussed.

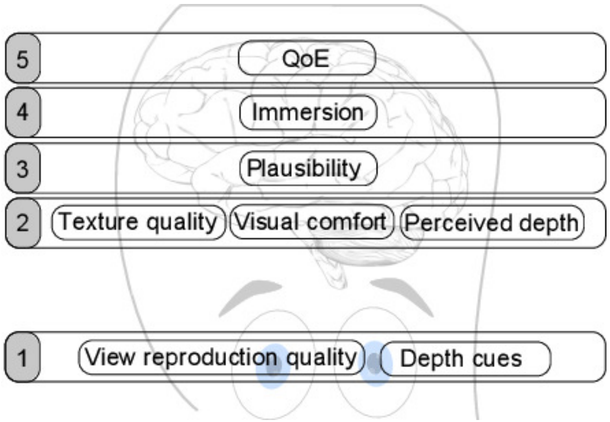


Figure 2.1: Multidimensional aspects of 3D QoE (Figure adapted from [8])

Different studies were performed to evaluate the individual factor and their relation. In the following, a list of studies with their respective evaluation criteria will be provided, to give an overview of what has been evaluated and how. The results themselves will then be discussed further in the following sections. The next subsection serves as an index for the next subsections. Table 2.1 provides a list of related work on the relationship between different 3D evaluation concepts which will be further discussed in the thesis.

| Ref. | What is addressed | Definitions | Test Method | What varied |
|---|---|---|---|---|
| [9] | Naturalness and visual experience as a function of blur and white noise | **Image quality:** excellence of the image<br>**Naturalness:** realistic or truthful reproduction of reality<br>**Depth percept:** amount of depth<br>**Viewing experience:** total experience related to the display | ACR (5 grade scale) | White noise and Gaussian blur level (4 levels each). |
| [3] | 3D Quality of Experience, Visual comfort as a function of coding and transmission conditions | **Quality of experience:** the overall experience<br>**Visual discomfort:** comfort compared to 2D viewing | ACR-HR (5 grade scale) | Different coding scheme: Simulcast, MVC, Frame Packing, 2D (4 levels each). Packet loss (2 levels for 2D and 3D Simulcast) |
| [4] | 3D Quality of Experience, Visual comfort as a function of coding and transmission conditions | **Quality of experience:** the overall experience<br>**Visual discomfort:** comfort compared to 2D viewing | ACR-HR (5 grade scale) | Simulcast, MVC (4 levels each), Frame rate reduction, downscaling (2 levels each), 2D. Packet loss short and long duration (on one view), different error concealments |
| [5] | Evaluation of Presence | **Presence:** the sense of being there<br>**Presence:** forgetting about the "real world" outside<br>**Presence:** something they had seen or visited | ACR<br>Continuous eval.<br>ruler-based method<br>verbal scaling<br>pair comparison | Review of methods. No tests. |
| [2] | Still images. Aymmetric and asymmetric JPEG compression. | **3D Image quality:** bad, poor, fair, good, excellent.<br>**Perceived depth:** numeric scale from 1 to 5<br>**Perceived sharpness:** numeric scale from 1 to 5<br>**Perceived eye-strain:** Rated on an annoyance scale. | ACR (5 grade scale) | JPEG compression (4 levels on each view. Full factors design). 3 different depth levels. |
| [10] | 3D videos. Image quality and effect of spatial and temporal resolution. Asymmetric conditions. | **Perceived quality:** bad, poor, fair, good, excellent. Rated by group 1.<br>**Perceived sharpness:** (same scale) Rated by group 2.<br>**Perceived depth:** (same scale) Rated by group 2. | DSCQS (Continuous scale 0-100) | Horizontal, vertical downscaling. Frame rate reduction. 2D and 3D. |
| [2] | 3D images. Viewing experience as a function of noise. | **Viewing experience:** Scale: bad, poor, fair, good, excellent.<br>**Naturalness:** Scale: bad, poor, fair, good, excellent | ACR (5 grade scale) | 6 levels of white noise. |
| [11] | 3D videos. Naturalness, Depth, Image Quality as a function of different video coding algorithms. | **Naturalness:** Scale: Bad, poor, fair, good, excellent.<br>**Depth:** scale: bad, poor, fair, good, excellent<br>**Image quality:** scale: bad, poor, fair, good, excellent | ACR (5 grade scale) | JM h.264 encoder, Simulcast, MVC (JMVC), JM and Side-by-side frame packing. |
| [12] | 3D videos. Visual experience as a function of Depth, Image Quality and Visual comfort. Content with no coding degradation. | **Visual comfort:** visual discomfort related to multisymptoms, e.g. eye strain, dry eyes, double vision.<br>**Depth:** amount of the perceived depth<br>**Image quality:** the quality of texture rendering, the level of visibility of visual artifacts and rendering details.<br>**Depth rendering:** the quality of the depth rendering depending on the subject's preference on the basic criteria related to stretching or compression of the reality and the shape of the objects<br>**Naturalness:** focuses on the evaluation of the natural appearance of images, i.e. whether the scene is more or less representative of reality<br>**Visual experience:** the overall quality of experience of the images in terms of immersion and the overall perceived quality. | SAMVIQ | Different shooting conditions. Image quality being only affected by geometrical distortions due to shooting condition. The different contents provide different ranges of comfort and depth quality as well. |

Table 2.1: Selected set of publications on 3D video QoE research and its relation with lower level factors. (ACR: Absolute Category Rating, ACR-HR: Absolute Category Rating with Hidden Reference, DSCQS: Double Stimulus Continuous Quality Scale, SAMVIQ: Subjective Assessment Methodology for Video Quality)

### *2.1.2 Image quality*

Image quality in the context of 3D images and videos has two different connotations: the *image quality* by itself and the *texture quality*. The distinction between the two terms is clear in the special case of Depth-based image coding and rendering where each of the stereoscopic images presented to the viewer are produced by a decoder using 2D images and depth information. In this specific case, it is necessary to make the distinction between the two terms: the *texture quality* is the quality of the 2D picture used by the decoder in addition to the depth information to synthesize a new image with a different viewing position. The quality of this synthesized picture will be characterized by the term *image quality*. Similarly, while capturing 3D video content using 3D cameras, distortions can appear on 2D images such as Barrel distortion, Pincushion distortion, color bleeding, ... e.g. due to the lens quality. Finally, coding may also affect the image quality. A complete review of distortions of 2D images in the context of 3D videos was performed by Boev et al [13]. In the general case, the *image quality* refers to the quality of the picture seen by observers which can be the result of depth-based image rendering while the *texture quality* refers to the quality of the pictures provided to the depth-based image rendering algorithms. In case of 3D stereoscopic video where only two stereoscopic views are considered and no depth-based image rendering algorithm is involved, the terms are similar.

Considering image quality the use of stereoscopic videos open new issues. In particular, the differences of quality between the two stereoscopic views and their perceived overall quality is an aspect which has been intensively studied. This relates to the binocular suppression theory [14]. Seuntiëns [2] studied this aspect in case of still images encoded using JPEG compression. First, the picture quality was found almost independent of the inter-camera base distance, therefore depth did not affect the image quality itself. Different quantizing parameters were applied to the left and right view following a full-factorial design. Results showed that high asymmetry in picture quality will strongly affect the overall quality, and pictures of lower quality as for a JPEG quantization parameter of 20 for each view can be rated higher than a strong asymmetry, for original for one view and a JPEG quantization parameter of 10 for the other view. In case of non-extreme asymmetric coding conditions, the overall perceived quality was found to be approximately equal to the average perceived quality of each individual view. However, these results were found to be degradation dependent, and as reported by Stelmach [10], in case of a degradation such as blurring, or equivalently downscaling, the overall quality of two stereoscopic views is driven by the highest quality view. Moreover, Stelmach [10], found that a downscaling ratio of half along both axes of one stereoscopic view, while the other is kept at full resolution, is not visible to the observers.

However, one issue still weakly studied in such kind of work is the long-term effect and resulting fatigue resulting from such approaches. The limited work on this topic is not due to a lack of interest, but rather due to the difficulty to address it, since it requires long tests per condition which makes it difficult to test many different conditions. Moreover, evaluating fatigue itself is a difficult task.

### *2.1.3 Depth*

The added value of 3D is to bring the perception of binocular depth to the overall user's experience. The overall perceived depth is resulting from the information of many different sources referred to as "depth cues". Depth within the context of 3D images and video has two different connotations: the *depth quantity* and the *depth quality*. *The depth quantity* relates to the amount of depth perceived in the 3D material, considering a specific setup. *The depth quality* is related to the plausibility of the depth rendering considering a specific setup including how the content was captured and rendered (see Figure 2.2). Both of these two aspects will be addressed more extensively respectively in Section 2.2 on foundations of depth perception and in Section 2.5 on technical implementations.

Seuntiëns [2] and Lambooij et al [9] studied the added value of 3D as compared to 2D. Test participants were asked to rate the naturalness of 2D and 3D images. White noise or blur was added to the images, and using the naturalness scores the added value of the 3D version of the image was compared to the 2D images by finding which amount of white noise added to the 3D presentation lead to similar ratings of the 3D and 2D images. Results show that the 3D

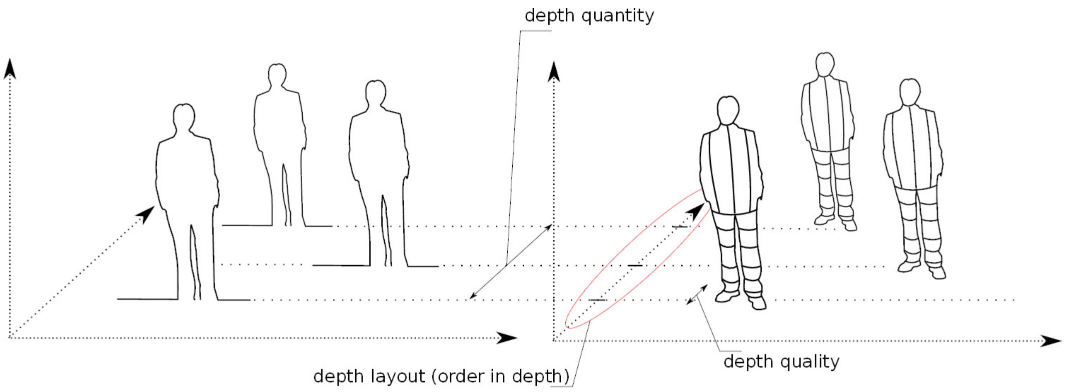


Figure 2.2: Different concepts related with depth perception: Depth quality, quantity, layout. The figure on the left side represents a lower depth quality than the figure on the right side.

effect was found to provide an improvement of naturalness equivalent to 2dB of white noise.
Yamagishi et al [11], used another type of distortion, namely video encoding using H.264. They showed that naturalness does not increase due to the presence of depth, and 2D and 3D were rated similarly. Such results may be explained by the context of the experiment, where test participants may have been too focused on the image quality degradation due to the video encoding, and naturalness is generally degraded by the artificial look of coding artifacts.
From these results it can be observed that evaluating 3D added value and its relation with texture quality is challenging, since it appears to highly depend on the context and employed method of the subjective evaluation. Therefore, applying distortions such as blur, white noise or degradation due to improper shooting conditions does not appear to drive the attention of the test participant away from the 3D effect as much as for coding. A possible interpretation of this result is the range of degradations on the image quality produced by these types of distortions. If highly distorted images are presented to the test participants, it appears that they will focus mainly on image quality rather than on the depth effect. As stated above, this effect may be increased or caused by the single stimulus test method applied in the reviewed studies.

### 2.1.4 Visual discomfort

A major issue regarding 3D Quality of Experience is *visual discomfort*. As described by Lambooij et al [15] *visual discomfort* is usually related to *visual fatigue*. *Visual fatigue* is related to “the decrease of performance of the human visual system”, and the *visual discomfort* is “the subjective counterpart of visual fatigue”. Hence, the visual fatigue relates to the long-term effect of visual discomfort. The sources of visual discomfort are multiple and will be addressed in Section 2.4. These include:

- Anomalies in binocular vision
- Geometrical distortions between the left and right images, crosstalk, binocular rivalries, color and brightness mismatch between views, suboptimal synchronization between views...
- Excessive binocular parallax
- Conflict between vergence and accommodation

The effect of visual discomfort on the overall QoE when watching 3D image and video material is strong and has been addressed in many different ways. For example, in [16, 17], the authors have looked into the effect of inter-camera distance and the effect on the overall QoE rating. Since the videos were presented without coding and transmission degradation, only perceived depth and comfort affected the scores. Results show that too high values of disparity result in a strong drop in Quality of Experience scores, and can be explained by the effect of high disparity values on visual discomfort.

Although visual discomfort is one of the key aspects of 3D, this thesis primarily addresses the perceived depth. The choice is motivated by the fact that the added value provided by the depth to the overall experience may be what would justify taking the “risk” of having discomfort. As a consequence, it is an aspect of strong interest.

### *2.1.5 Models of QoE*

Different models of QoE have been proposed in the literature, see e.g. Seuntiëns [2], in the following section, some of these models will briefly be reviewed in relation to 3D video QoE. The overall 3D Quality of Experience (level 4 in Figure 2.1) was expressed as a function of Naturalness (level 3 in Figure 2.1) and visual comfort (level 4 in Figure 2.1). The naturalness was expressed as depending on the image quality and the depth (see Figure 2.3). Similarly, the model from Chen [12] expressed the overall QoE as a linear combination of 2D image quality, depth quantity and visual comfort (see Figure 2.4). This model is different from the model by Seuntiëns by the aspects it addresses. Indeed, in the model from Chen, the 2D image quality does not refers to quality degradations such as coding or noise, but the quality of the stereoscopic images due to the shooting condition. The second difference concerns the depth, which is decomposed into two aspects: depth quality and quantity. As a consequence, image quality coupled to the depth quantity enables to address issues such as the quality of the depth rendering. Similarly to Seuntiëns, naturalness is a function of the image quality and visual comfort and depth instead of being expressed as a single notion is decomposed into its two aspects: depth quality and depth quantity. The overall visual experience is modeled by a combination of the different factors.

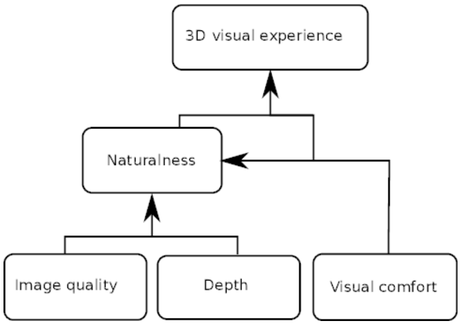


Figure 2.3: Model of QoE from Seuntiëns [2]

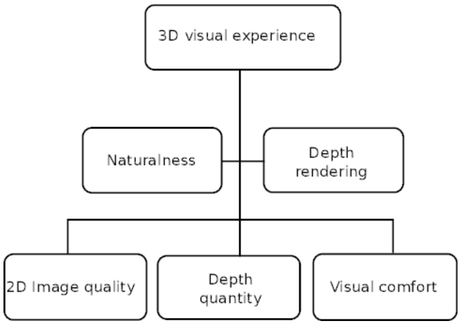


Figure 2.4: Model of QoE from Chen [12]

The models described up to now relate the interaction between *depth*, *image quality* and *visual comfort* with the overall 3D experience. However, the perception of these aspects does not only depend on the 3D video itself, but also of the equipment such as the display and its rendering abilities. To address this Engeldrum [18] defined a model for image quality which includes this technology variable directly into the image quality model. Figure 2.5 illustrates the proposed model. It can be seen that there is a relationship between the technology variable (how the image is rendered), how it is evaluated, and the final customer’s rating. The image quality circle is closed by linking the rating to the technology variable. Indeed, technology affects the perception of the image quality due to the context of use, and expectations with regard to previous experiences with the device. The model proposed by Engeldrum only restricts to image quality, therefore Lambooij [9] has extended it to the other factors involved in 3D QoE (see Figure 2.6). The other aspects, *depth*, and *visual discomfort* are also involved in a similar circle as the image quality circle described by Engeldrum. Then, similarly to Seuntiëns, the overall 3D experience can be expressed as a combination of the three factors *image quality*, *depth quality* and *visual comfort*. To further study the relationship between image quality and depth and their integration into a 3D Quality model, Lambooij [9] designed two different models to also explain the evaluation concepts *naturalness* and *visual experience* as a linear combination of image and depth quality. A generic schema of the integration model is shown in Figure 2.7. Each evaluation concept (EC) is expressed as a linear combination of image quality (IQ) and perceived depth (D) (Eq. 2.1).

$$EC = \alpha \cdot IQ + \beta \cdot D \tag{2.1}$$

To determine the relationship between these factors and the concept *naturalness* and *visual experience*, different experiments were conducted with different conditions of white noise and blur, different kinds of displays and contents. Curve-fitting led to a weighting of 0.74 and 0.26 respectively for $\alpha$, and $\beta$ when naturalness is modeled. Similarly weighting of 0.82 and 0.18 were found for $\alpha$, and $\beta$ when the visual experience is targeted.

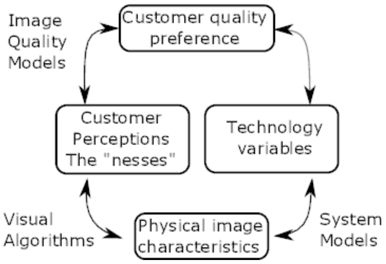


Image Quality Models
Customer quality preference
Customer Perceptions The "nesses"
Technology variables
Visual Algorithms
Physical image characteristics
System Models

Figure 2.5: Model of image quality proposed by Engeldrum [18]

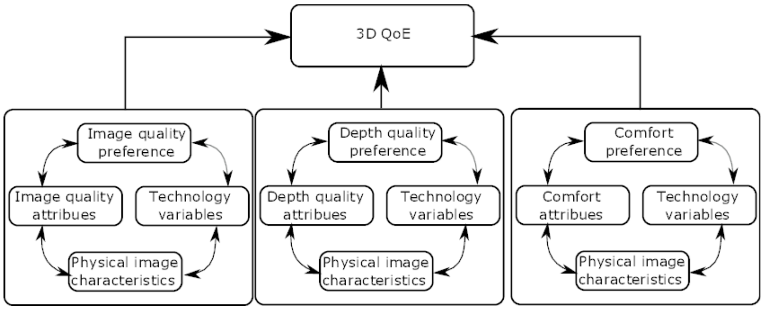


Figure 2.6: Model of QoE from Lambooij (Figure from [19])

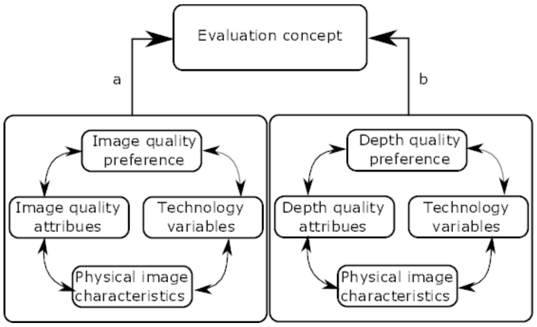


Figure 2.7: Model of 3D Quality for evaluating the Naturalness and the Visual Experience (Figure from [9])

### 2.1.6 *Interaction between depth and other factors*

In this work, it is the *perceived depth* which will be addressed, since it provides the added value of 3D compared to 2D videos. Depth is not independent of the other factors presented in the previous section: e.g. image quality and visual comfort. In this section, results of the literature on interactions between depth and the others mentioned factors will be presented.

#### 2.1.6.1 Perceived depth and image quality

The amount of perceived depth and its relation with image quality have been investigated in a number of studies. Seuntiëns [2] studied the effect of JPEG compression of still images on the perceived depth, and no clear effect could be seen on the depth scores (Figure 2.8). Two kinds of degradation were considered by Lambooij et al [9], Gaussian noise and Gaussian blur. The effect of Gaussian noise on the perceived depth was rather small, but still significant and Gaussian blur was found to have a high effect on the depth scores. These results are explained by the fact that edges contribute considerably to depth perception as reported by Bülthoff [20]. Considering that Gaussian noise has a limited effect on the edges and Gaussian blur strongly affects the edges, it was expected to see such result. Yamagishi et al [11] also found that in the case of video encoded at different bitrates image quality affects depth, and at too low quality the depth effect is not perceivable anymore. Figure 2.9, depicts their results. In their experiment, different source sequences were encoded at different bitrate and with different coding schemes. The coding schemes were side-by-side using the full resolution of the source sequences and were encoded with H.264, side-by-side with half of the horizontal resolution and encoded with H.264, and multi-view coding (MVC) which takes into account the inter-frame redundancy. The different processed video were then evaluated on two evaluation concepts: perceived depth and image quality. It can be seen that the relation between bitrate and perceived depth was found to be content-dependent, which may be due to the complexity of the contents themselves, resulting in different image quality for a given bitrate, so that the depth was affected differently.

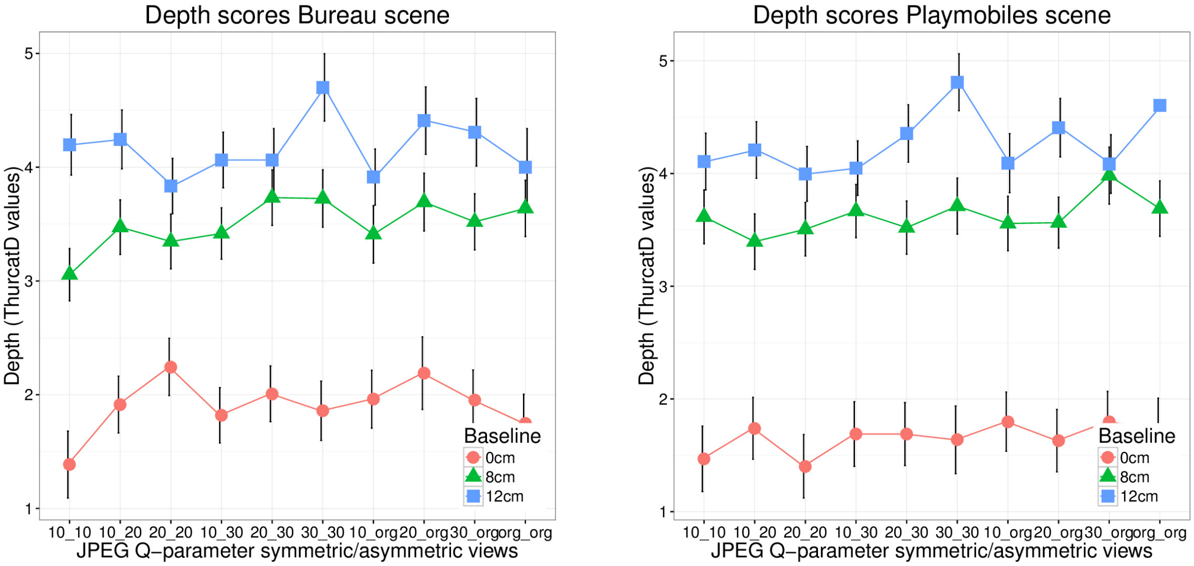
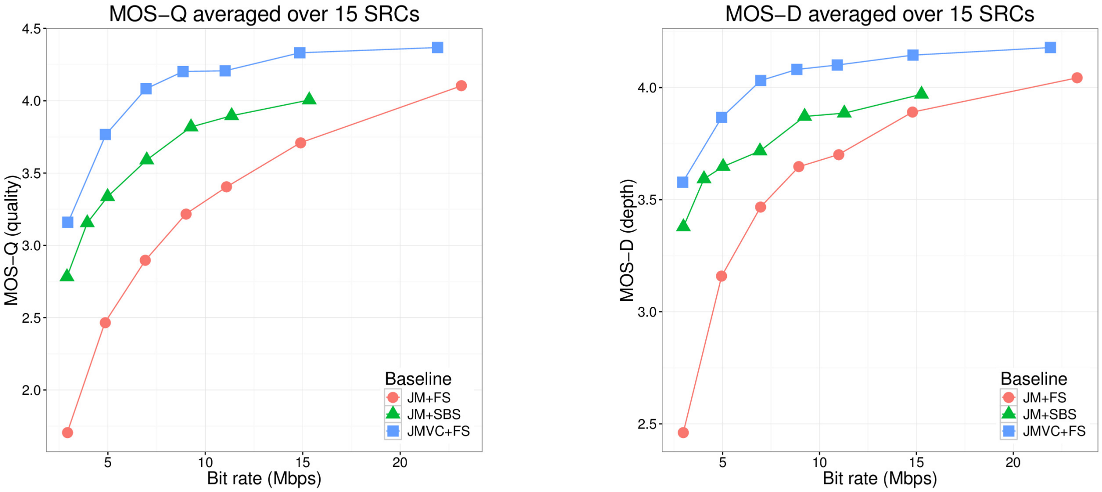


Figure 2.8: Relation between depth quantity and image quality, results from Seuntiëns [2]. B: the baseline between the two cameras. On the horizontal axis different combination of quantization for left and right view are provided. For example 10_20 means a quantization parameter of 10 for the left view, and 20 on the right view.

.

### *2.1.7 Summary*

In this section the problem of modeling 3D Quality of Experience (QoE) based on different factors such as *image quality*, *depth* and *visual comfort* was addressed. Different general models from the literature for predicting QoE were presented. The modeling of high-level evaluation concepts such as *naturalness*, and *visual experience* have been


MOS−Q averaged over 15 SRCs

MOS−D averaged over 15 SRCs

Figure 2.9: Relation between depth, image quality, and video compression using different coding schemes. The left figure illustrates the relation between image quality and bitrate for the different coding schemes. The right figure addresses the relation between bitrate and perceived depth quantity. Figure reproduced from Yamagishi [11].

discussed in case of very specific degradations: blur, or white noise. As shown in this section, depth was found to be one factor involved in the overall QoE formation, and thus has to be investigated in detail. A more detailed description of depth perception will be provided in the next section.

## 2.2 Human perception of depth

In this section an overview of the literature on depth perception in 3D images and video sequences will be provided. The goal is to enable an overall understanding of how depth is perceived based on the different kinds of information available to the eyes of a person watching the world around her. There has been a large body of research performed to analyze how the different sources of information contribute to the perception of depth.

Different factors underlie the notion of depth perception. At first, it is necessary to identify and define these, and to specify which ones are targeted.

### *2.2.1 Different factors*

Regarding the perception of depth, one can talk about the "depth quantity", "depth quality", or the "scene layout" (see Figure 2.2). Each of these factors describe a different aspect of depth perception. The first one, the depth quantity, describes how much depth can be perceived in the scene based on all the depth cues available. The "depth quality" is a factor involved in case of 3D rendering on any kind of displays. In this case, two steps are involved: capture or production of a scene and its rendering. Depending on production and display conditions, the geometry of the rendered 3D objects may be affected, and distortions of their shape can appear [21, 22, 23, 24]. One extreme case of such distortion is the "cardboard effect" where the objects appear flat in different depth planes [13]. Here, depth quality addresses how well depth is presented to viewers, in terms of their "depth quality" perception. The last factor is the scene layout. It characterizes the ability to order the objects in depth [25]. In the following of this thesis, except where explicitly stated, it is the depth quantity which is targeted.

### *2.2.2 Depth cues*

There are two different categories of depth cues, the binoculars and monocular depth cues. Binocular depth cues result from the retinal images in the two eyes of the observer. Figure 2.10 depicts two examples of binocular depth cues. The schema on the left side represents the retinal binocular disparity, or stereopsis: the two eyes see two distinct objects and the projection of these objects on the retina appear to be at different locations on the retina of each eye. If one point is the point of *fixation*, the differences between the position of theses two projections are the *retinal disparity*. This information is processed by the brain to estimate the relative position in depth between the two objects, and is the most important binocular depth cue. The second binocular depth cue depicted in Figure 2.10 is the vergence. The two eyes converge on the object under study. The information from the orientation of each eye is an indication of absolute position in depth, provided to the brain.

The second category of depth cue is the monocular depth cues. Figure 2.11 depicts some of them. More detailed explanations follow in section 2.2.2.2. All monocular and binocular depth cues have different abilities for the characterization of depth. Some can provide absolute position in depth of an object such as the vergence, the motion parallax, the relative size. Some can only provide relative position in depth such as the defocus blur, the binocular disparity, the interposition. In addition, Cutting and Vishton showed that the distance of observation is also of high importance and studied the threshold of difference of depth depending on the viewing distance [25] (See Figure 2.12). Results show that some depth cues are always discriminatory, such as occlusion or relative size, and others like binocular disparities or motion parallax are only informative within a certain depth range. For example, binocular disparities can be used between 0-17 m and motion parallax between 0-1000 m. All of these depth cues have different reliability, different discriminative power based on the viewing distance and can interact with each other in the process of the construction of the perception of depth.

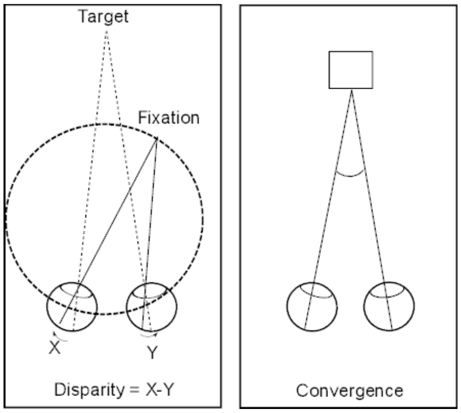


Figure 2.10: Binocular depth cues.

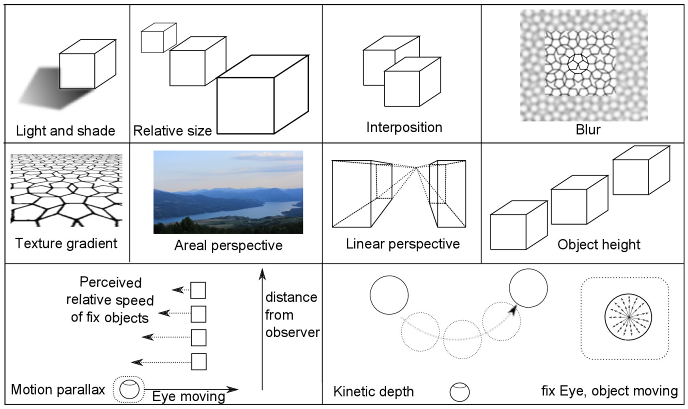


Figure 2.11: Monocular depth cues.

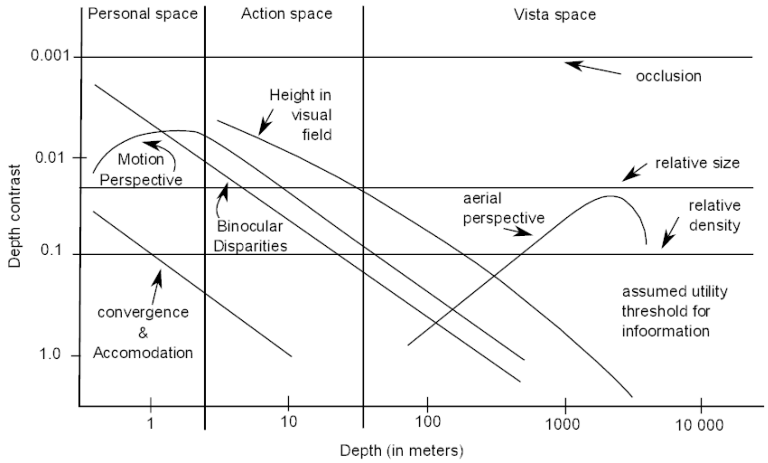


Figure 2.12: Depth contrast perception (e.g. ability to perceive differences of depth) as a function of the viewing distance. Results and Figure from Cutting & Vishton [25].

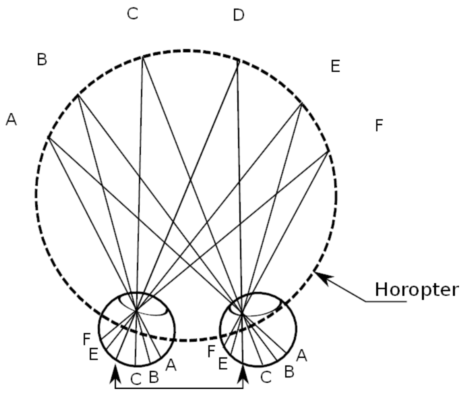


Figure 2.13: Horopter: all the different points belonging to the Horopter will appear at the same location on the retina.

#### 2.2.2.1 Binocular depth cues and depth perception

This subsection focuses on the particular case of binocular depth cues. Figure 2.13 depicts the horopter, the points located on it will be perceived at the same location of the retina. In the space described by the horopter (Figure 2.10 left), the objects in front of the horopter have negative disparities. They are also called crossed disparities. And vice versa, the points located beyond the horopter have positive or uncrossed retinal disparities.

It is possible to *fuse* two stereoscopic images, and then be able to perceive a single image from the two retinal images seen by each eye, if the disparities belong to a limited area. Outside of this area, persons suffer from *diplopia* which corresponds in seeing double. The area where the stereo vision is possible is called the *Panum's fusional area*, and depends on the eccentricity from the fovea: on the fovea the retinal disparities should be limited to 0.1° to ensure binocular fusion, but with higher eccentricities such as 6°, the range of binocular disparities can increases to 0.33°, and at an eccentricity of 12° it can reach 0.66°[15].
One of the different factors which contribute to depth perception is illustrated in Figure 2.10, the distance between the two eyes: the inter-pupillary distance, which strongly affects the amount of retinal disparities. Studies have shown that the majority of adults have an inter-pupillary distance in the range from 50 to 70 mm, with a mean and median of

63 mm [26]. These values are used as an average setting for both depth perception and visual comfort studies.
As explained previously, the area where fusion is possible is particularly small. Without movement of the eyes (vergence movement) and for short duration stimuli, fusion limits of 27 arcmin for crossed disparities and 24 arcmin for uncrossed disparities were found. As reported by Lambooij [15], "many factors have been found having an effect on fusion. These include eye movements, stimulus properties, temporal modulation of retinal disparity, exposure duration, amount of illuminance, and individual differences". The limit of fusion was found to decrease with small, detailed and stationary objects and increase with larger, moving objects and in case of objects existing in the periphery of fixed objects [27, 28, 29, 30]. "With longer duration and vergence movement, retinal disparities can be as high as 4.93° for crossed disparities and 1.57° for uncrossed disparities before producing diplopia" (As reported by Lambooij [15]).
In addition to retinal binocular fusion, or Stereopsis, another cue from the stereoscopic vision is the *convergence*. Indeed, when looking at an object the two eyes converge on the object under observation as depicted in Figure 2.10 (right). The angle of convergence provides an absolute measurement of distance between the object and the observer location. As reported by Cutting & Vishton [25] and depicted in Figure 2.12, this depth cue is mainly effective for distances less than 10 meters.

#### 2.2.2.2 Monocular depth cues and depth perception

In addition to binocular depth cues, monocular cues also contribute to the perception of the depth in images. The monocular depth cues include:

***Motion Parallax:*** "when an observer moves, the apparent relative motion of several stationary objects against a background gives hints about their relative distance. If information about the direction and velocity of movement is known, motion parallax can provide absolute depth information", as reported by Ferris [31, 32] (See Figure 2.14).

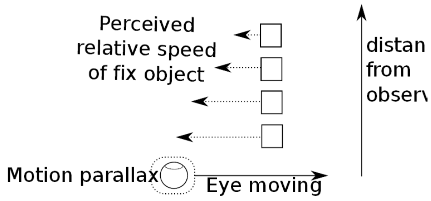


Figure 2.14: Motion parallax

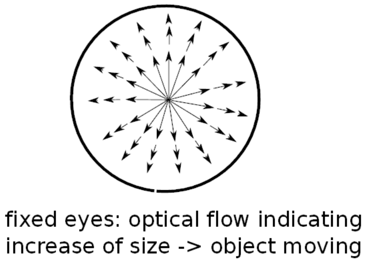

fixed eyes: optical flow indicating
increase of size -> object moving

Figure 2.15: Depth from motion

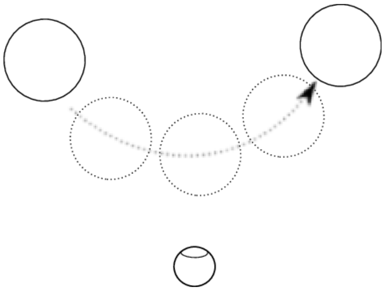


Figure 2.16: Kinetic depth effect

***Depth from motion:*** "When an object moves towards the observer, the retinal projection of an object expands over a period of time, which leads to the perception of movement in a line towards the observer. Another name for this phenomenon is depth from optical expansion", as reported by Swanston [33, 32]. (See Figure 2.15)

***Kinetic depth effect:*** "If a stationary rigid figure (for example, a wire cube) is placed in front of a point source of light so that its shadow falls on a translucent screen, an observer on the other side of the screen will see a two-dimensional pattern of lines. But if the cube rotates, the visual system will extract the necessary information for perception of the third dimension from the movements of the lines, and a cube is seen", as reported by William [34, 32]. (See Figure 2.16)

***Linear perspective:*** The term linear perspective is used since the 14th century. A definition can be found in Webster's dictionary as "the technique or process of representing on a plane or curved surface the spatial relation of objects as they might appear to the eye; specifically : representation in a drawing or painting of parallel lines as converging in

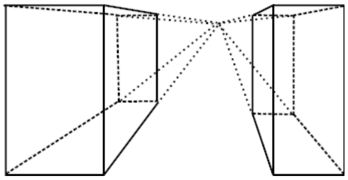


Figure 2.17: Linear perspective

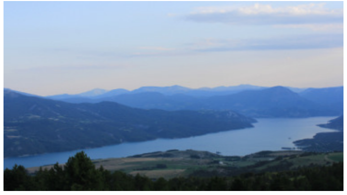


Figure 2.18: Aerial perspective

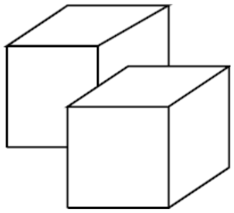


Figure 2.19: Interposition

order to give the illusion of depth and distance" [35]. (See Figure 2.17)

***Relative size:*** Cutting defines the Relative Size as "a measure of the angular extent of the retinal projection of two or more similar objects or textures" [36]. It "arises from the differences in the projected angular sizes of two objects that have identical sizes and are located at different distances. If the assumption that the two objects have identical physical sizes is met, then from the ratio of their angular sizes, it is possible to determine the inverse ratio of their distances to the observer. In this way, metrically scaled relative-depth information can be specified" as reported by Weiner [37]. (See Figure 2.20)

***Familiar size:*** As described by Weiner [37], if the size of the object is known, the angular size of the object can be used to determine the absolute distance information.

***Absolute size:*** Even if the actual size of the object is unknown and only one object is visible, a smaller object seems further away than a large object that is presented at the same location, c.f. Sousa et al. [38, 32].

***Aerial perspective:*** Cutting defines the Areal perspective as follows: "refers to the increasing indistinctness of objects with distance, determined by moisture and/or pollutants in the atmosphere between the observer and these objects. Its perceptual effect is a decrease in contrast with distance, converging to the color of the atmosphere" [36]. (See Figure 2.18)

***Accommodation:*** This is an oculomotor cue for depth perception. Weiner defines it by referring "to the change in shape of the lens that the eye performs to keep objects at different distances in focus. Changes in accommodation occur between the nearest and the farthest points that can be placed in focus by the thickening and thinning of the lens" [37]. This can be used as an information about the depth of the object under observation.

***Interposition:*** Occlusion (also referred to as the interposition) "occurs when an object partially hides another object from view, thus providing ordinal information: The occluding object is perceived as closer and the occluded object as farther" (Weiner [37]). This information only allows the observer to create a "ranking" of the object positions. (See Figure 2.19)

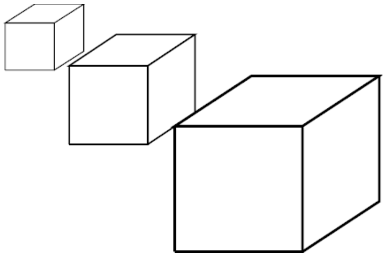


Figure 2.20: Relative size

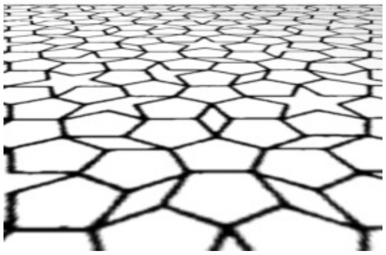


Figure 2.21: Texture gradient

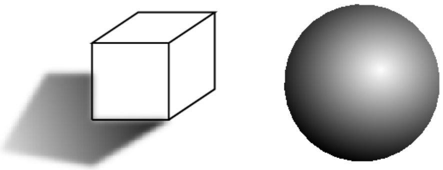


Figure 2.22: Light and shade

***Texture gradient:*** Texture can provide information about the location of the objects. Cutting and Millard divided texture gradient into three categories: *perspective*, *compression* and *density* [39]. *Perspective*: due to the linear perspective, the size of the objects decreases with the distance to the observer, therefore the size of the individual texture element (texel) is affected by this phenomenon. *Compression*: relates to the ratio between the width and height of the texels. The aspect ratio of the texels will be affected by the position. Finally, *density* refers to the spatial distribution of the texels in the image. These different aspects are illustrated in Figure 2.21

***Light and shade:*** Weiner defines this factor as referring "to the smooth variation in image luminance determined by a combination of three variables: the illuminant direction, the surface's orientation, and the surface's reflective properties". Depth from shading is complex since different conditions can result in similar shadings. However, if the illuminant is known it is possible to derive information about the structure.(See Figure 2.22)

***Defocus blur:*** Related to accommodation, objects in focus are perceived as sharp and objects located at far distances from the objects in focus appear as blurred. The blur provides an absolute value of the relative position in depth of the objects. Indeed, from this cue alone it is not possible to distinguish between objects located in front or in the back of the object in focus.

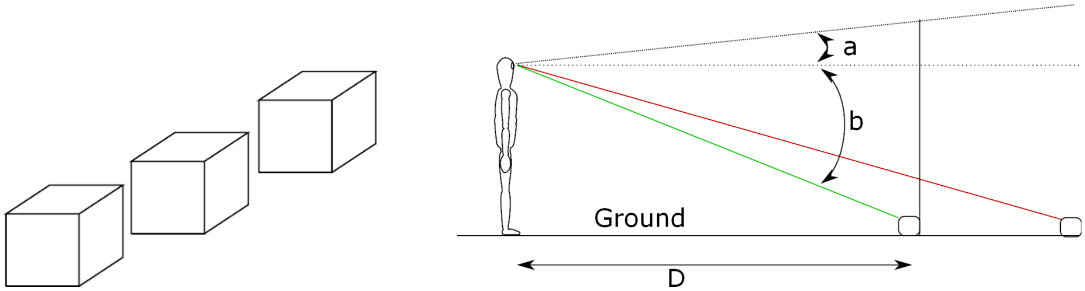


Figure 2.23: Object height. The position in height of the objects compared to the horizon line can be used to derive the distance of the objects from the observer.

***Height in the Visual Field and the Horizon Ratio:*** Considering an object and an observer located on the same ground, the farther the object is from the observer, the closer the object is to the horizon line. Therefore, the vertical height of the object can be used as an information of the position in depth, (see Figure 2.23, left) Sedgwick defined the geometry behind this phenomenon [40] which can be found in Equation 2.2 (See Figure 2.23, right).

$$h = \frac{tan\ b + tan\ a}{tan\ b} \tag{2.2}$$

#### 2.2.2.3 Discussion about monocular depth cues

All the different monocular depth cues are not necessarily orthogonal, and some depth cues can be perceived as a combination of different other depth cues. In their work on depth threshold perception, Cutting and Vishton [25] discarded several depth cues in their study due to these strong interdependencies (see Figure 2.12). The depth cues omitted are:

- *Texture gradient:* As stated in the previous section, texture gradient can be characterized along three different axes: gradient size, density, and compression. The gradient size depends on the difference between the maximum and minimum size of the texture element. Therefore this component relates to the relative size cue. The density can be considered as a depth cue itself. And finally, compression was found by Cutting and Millard to have a limited effect on the visual system to reveal depth [39].

- *Linear perspective:* linear perspective was analyzed by Elkins [41] as a systematic combination of several sources: texture gradient (size, density, and compression) and occlusion. The convergence of the parallel lines could be expressed by means of object size, density and compression as reported by Taylor [42].
- *Brightness, light and shade:* Cutting and Vishton discussed that shadings are the main source of depth and not brightness. Shadows provide information about the shape of the object which can be considered as an application of transparency. Therefore, it can be related to the aerial perspective. A second motivation of Cutting and Vishton for not considering it is that shading will provide information on the shape of the object rather than its position compared to other objects.
- *Kinetic occlusion and disoclusion:* kinetic occlusion and disoclusion can also be considered as a specific case of occlusion where no luminance contrast at occluding edge is required. This can also be related to the motion parallax (Cutting [36]).
- *Gravity:* Gravity can also be seen as a source of depth information by considering the acceleration of dropped objects or thrown objects which will vary with the viewing distance, as studied by Watson [43]. This effect is related to motion parallax which results from the perception of speed of objects at different distances.

This shows that all the different cues can interact between each other because of the definition of the depth cues by itself or their underlying physical or geometrical foundations. Further type of interaction between the depth cues will be addressed in the next subsection.

#### 2.2.2.4 Interactions between depth cues

The possible type of interactions between and combinations of depth cues has been categorized by Bülthoff and Mallot into five categories [44]. These categories are:

- *Accumulation:* This corresponds to the pooling of depth cues in the final stage of the depth estimation process. The depth cues are considered after taking into account any type of interaction they could have between each other. These interactions may have resulted in enhancing or decreasing their contribution to the overall depth perception. The concept behind accumulation is that a judgment based on cues which agree will be more reliable than a judgment based on one cue. Moreover, when depth cues conflict with each other, the resulting depth perception may correspond to an intermediate value between the different depth cues. Based on these considerations, the most efficient way to combine depth cues is a weighted mean with weights depending on the reliability and efficiency of each cue [45]. However, in some cases an average may not be an appropriate pooling strategy due to underestimation of the perceived depth from some cues. In this case, an additive strategy may be a better approach (examples will follow in the next section). Therefore, there is the need to define a function which will perform a tradeoff between the two pooling strategies: summation or averaging.

- *Veto:* The perception will be based only on one cue, which forces its value on the other depth cues.

- *Cooperation:* The cooperation, in contrast to *accumulation*, appears in the early stage of the depth cue pooling and relates to the interactions between depth cues: the fact that one depth cue will enhance or decrease the contribution of another depth cue to the overall perception.

- *Disambiguation:* This is additional information provided by one cue about another depth cue to solve ambiguous cases. For example, defocus blur can only provide an absolute value of difference in depth: it cannot distinguish between positive and negative change of depth around the focus plane. Another depth cues such as binocular disparity or other monocular depth cues can then help to solve this ambiguity.

- *Hierarchy:* The concept of the hierarchy is based on the fact that information from several depth cues can be considered as raw data for another depth cue.

Beyond the categorization of Bülthoff and Mallot, two further types of interaction between depth cues were defined:

- *Cue dominance [46]:* This phenomenon appears when two cues contradict each other, in this particular case one cue may dominate over the other one. This is similar to *Veto* except that the term *Veto* is used for small contradiction between the depth cues.
- *Cue promotion [47]:* The cue promotion is a scaling which is performed to align different depth cues using parameters extracted from other depth cues.

## *2.2.3 Models for depth cue fusion*

Based on the different sensory input, several models have been defined in the literature. These models have been classified into two categories by Clark and Yuille: the weak fusions and the strong fusions [48].

### 2.2.3.1 Weak fusion

Most of the studies have considered weak fusion algorithms. These models assume no interaction between the different depth cues, and the depth cue pooling can be limited to the "accumulation". Depth from individual depth cues is computed separately, then a weighted linear summation of each estimated depth value from each cue is performed by weighting the contribution of each cue based on its reliability. This type of model has the advantage to be simple and modular since every depth cue can be analyzed individually. However, there are several limits to these models which were explained by Landy et al. [49]: in addition to the fact that any interaction between depth cues is neglected, a strong limitation is to put every individual depth cue to the same scale, since some may be expressed in physical units such as meters, and others may be in other units. Moreover, some depth cues such as the motion cues may not always be available, and having these cues out of the pooling equation is different than setting their contributions to zero in a linear equation. Besides, the weights based on the confidence in each metric is strongly dependent on the characteristics of the considered signal. This requires the weight to be dynamic which is then difficult to define. All these aspects make the weak fusion model applicable only in certain cases.

### 2.2.3.2 Strong fusion

Alternatively, the second main category of fusions is that of strong fusion. These models assume strong interactions between the different depth cues. The depth cues cannot be computed separately and depend on each other for defining how they contribute to the overall depth perception. One implementation of this paradigm as reported by Landy et al. [49] is provided by Nakayama and Shimojo [50], where the perceived depth in a scene is based only on one of the different depth cues. The selection of the cue used for the perception of depth is made by choosing the most plausible depth prediction between the different predicted depth values obtained from each cue. This selection is performed using information on the scene under study.

### 2.2.3.3 Hybrid approaches

Intermediate models between the strict weak and strong fusion have been proposed. For example, the model defined by Landy et al. [49], called "Modified Weak Fusion" (MWF), is restricted to linear combinations of independent depth cues. This is a weak fusion approach, but allows only one particular case of interaction between depth cues: the cue promotion. This interaction is part of the strong fusion approach making the model an intermediate approach between the two strict definitions of weak and strong fusion.

| | | | Depth cues analyzed | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Author | Year | Ref. | B | L | T | D | R | I | S | M | K | H | Other | Fusion |
| Dosher et al | 1986 | [51] | X | X | | | | | X | | | | | Linear |
| Bruno and Cutting | 1988 | [52] | | | | | X | X | | X | | X | familiar size | Additive |
| Johnston et al | 1993 | [53] | X | | X | | | | | | | | | Linear / based on viewing distance |
| Buckley and Frisby | 1993 | [54] | X | ∼ | X | | | | | | | | Outline cues | |
| Landy et al | 1991 | [55] | | | X | | | | | | X | | | Linear / based on cue reliability |
| Rogers and Collett | 1989 | [56] | X | | | | | | | X | | | | Linear / based on reliability |
| Wang et al | 2011 | [57] | X | | | X | | | | | | | | Linear / based on viewing distance |
| Held et al | 2012 | [58] | X | | | X | | | | | | | | Linear / based on viewing distance |
| Ernst and Banks | 2002 | [59] | X | | | | | | | | | | Haptic | MLE |
| Hillis et al. | 2004 | [60] | X | | X | | | | | | | | | MLE |
| Lovell et al. | 2012 | [61] | X | | | | | | X | | | | | MLE |
| Massaro | 1988 | [62] | | | | | X | X | | X | | X | familiar size | Multiplicative (FLMP) |
| Landy et al | 1995 | [49] | X | | X | | | | X | | X | | | Linear with cue promotion |
| Bülthoff and Mallot | 1988 | [20] | X | | | | | | X | | | | edges | Cue veto |
| Ogle | 1938 | [63] | X | | | | X | | | | | | | Cue veto |
| Braunstein et al | 1982 | [64] | | | | | | X | | | X | | | Disambiguation |
| Blake and Bülthoff | 1991 | [65] | X | | | | | | | | | | Specularitites | Disambiguation |
| Yuille and Bülthoff | 1995 | [66] | X | | X | | | | X | X | | | | Bayesian decision model |
| Girshick and Banks | 2009 | [67] | X | | X | | | | | | | | | Bayesian model |
| Nakayama and Shimojo | 1992 | [65] | X | | | | | | | | | | | Bayesian model |
| van Ee et al | 2003 | [68] | X | X | | | | | | | | | | Bayesian model |

Table 2.2: Subjective experiments on depth cue combinations. *B*: Binocular depth, *L*: Linear perspective, *T*: Texture gradient, *D*: Defocus blur, *R*: Relative size, *I*: Interposition, *S*: Light and shades, *M*: Motion parallax, *K*: Kinetic depth, *H*: Height

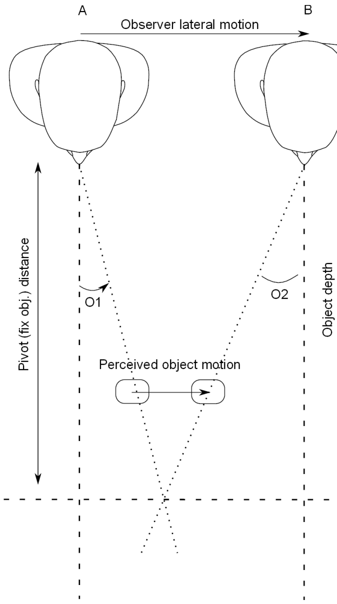


Figure 2.24: Evaluation based on parallax

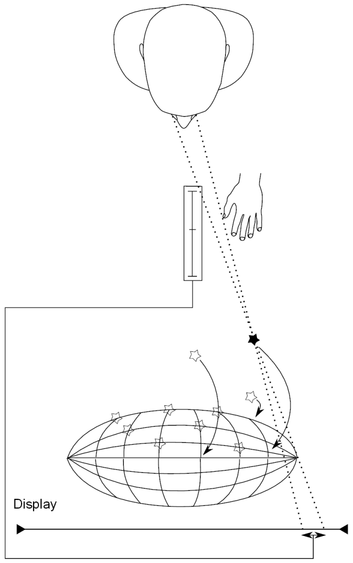


Figure 2.25: Evaluation using test spots

## 2.3 Results on depth modeling

In this section, results supporting the different models and how depth cues combine with each other will be presented. The experiments, the depth cues studied, and types of models used are categorized in Table 2.2 and will be discussed in Subsection 2.3.2.

### *2.3.1 Subjective depth evaluation methods*

In order to be able to study the combination of depth cues into a prediction of the overall depth perception, accurate methods for subjective evaluation are required. This subsection will present different measurement methods reported in the literature. A first method illustrated in Figure 2.24 is the one proposed by Gogel [69]. The idea behind the method is to evaluate the perceived position in depth in an indirect manner using parallax cues. Indeed, if the observer changes his position laterally the objects will appear to shift laterally as well. The shift is proportional to their distance in depth to the screen and the direction of the shift will also be dependent on their position relative to the display: the shift will be in the same direction as the observer's motion if the object pops out of the display and will be in the opposite direction if the object is in the zone within the display. By measuring the angles *O1*, *O2* reported by the participant, the displacement, and the viewing distance it is possible to have a measurement of the position in depth of the objects.

An alternative illustrated in Figure 2.25 is proposed by Bülthoff and Mallot [20]. The idea of this methodology is to ask the test participants to align several dots on a 3D image. This 3D image is the result of the combination of different monocular and binocular depth cues. Through the dots, the test participant can only adapt the binocular disparity and therefore, define an equivalency between binocular depth cues and the combination of several depth cues contained in the image (binocular disparity and shadings in [20]). A derived approach consists in adapting an ellipsoid defined using binocular depth cues to make it correspond to another ellipsoid defined by both monocular and binocular depth cues. Another alternative illustrated in Figure 2.26 is described by Johnston [70]. Similarly to the quality ruler [71], it consists in asking test participants to evaluate the test signal on a well-controlled scale of other stimuli. Participants should then select one of the reference stimuli, which best corresponds to the stimulus under

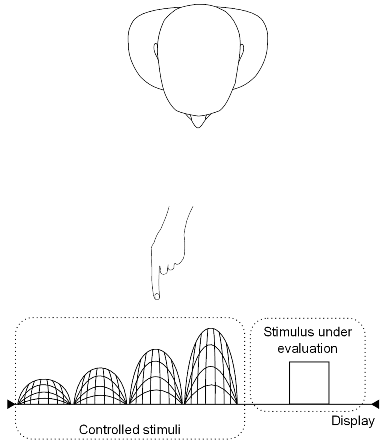


Figure 2.26: Evaluation using controlled stimuli

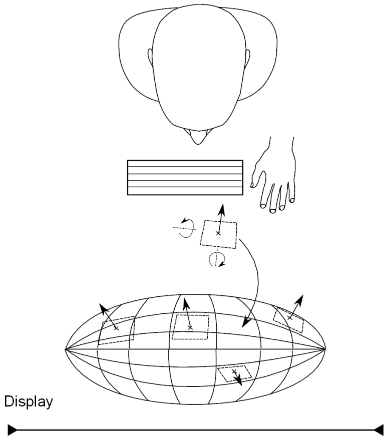


Figure 2.27: Evaluation by defining surface normals

evaluation. In the particular case of [70], the reference stimuli are several cylinders having continuous curved surfaces and having different ranges of depth. The test stimuli are different patterns of randot stereograms.

Other alternatives consist in comparing two different stimuli selecting which presentation shows the expected property. This property can be the largest expansion in depth as in Landy [49], or which is the tallest as in Ernst et al [59], or which has the biggest slope as in Girshick et al. [67], or which is the orientation of the rotation: clockwise or counter clockwise as in Braunstein et al. [64]. Alternatively, to decrease the number of paired-wise comparisons, it is also possible to compare two pairs of two stimuli and ask to report a property regarding the pairs themselves. For example, the largest depth expands.

Another approach is to let the test participants describe the orientation of a surface by its normal vectors. This method is illustrated in Figure 2.27, and was studied by Stevens and Brook [72]. The concept behind this methodology is to ask test participants to define the normal vectors of the surface under study. Alternatively, Van Ee [68] let the test participants adjust the orientation of lines to describe the slant of the stimulus (a 3D plane).

In the work of Dosher et al. [51] the task was different, and observers were confronted with a forced choice between two alternative options to explain their understanding of the scene. Participants were asked to report if what they saw is a cube or a truncated pyramid, or what the first direction of the rotation of the stimulus was: left or right.

Another alternative giving more freedom to the test participant is proposed by Tittle and Braunstein [73], who asked the observers to report the depth-to-height ratio of the stimuli under study.

Rogers and Graham [74], chose to gradually increase the amount of binocular disparities and let the test participants notify from which amount of disparity they can perceive the depth in the proposed stimulus. In this case, the stimulus under study is randot dots describing an oscillation in depth.

Finally, another approach can be found in the literature as presented by Bruno and Cutting [52] which consists of directly asking the test participants to evaluate the perceived depth between objects on a rating scale. In the particular case of [52] a scale from 0 to 100 was used, 0 meaning no distance between objects and 100 being the “maximal exocentric separation”.

### *2.3.2 Evidences to support models*

After having presented the evaluation methods in the previous section, the following section describes studies which have been conducted and reveals the different types of interactions possible between depth cues and the respective models.

#### 2.3.2.1 Weak fusion

Weak fusion models have been widely used, reflecting the fact that this kind of model performs well in many cases. Dosher et al [51] used a linear model between binocular depth, perspective and luminance to model the ability of an observer to distinguish kinetic, based on a viewing tests on the ability of the observer to see the direction of a rotation. In this study, stereopsis was found to be the dominant cue in static presentation, and it was dominant in most of the dynamic presentations as well. Another result is a recency effect which was observed: when a static presentation precedes a dynamic one, the static presentation strongly influenced the dynamic one.

To model the interaction between other depth cues, a linear model was used by Bruno and Cutting [52], where motion parallax, occlusion, height, and familiar size were considered. The underlying test used was to ask the test participants to rate the magnitude of exocentric distances using a forced choice between two stimuli. The outcome suggests that depth cues appear to be additive, independent, and each cue was referred to as “minimodules”.

Johnston et al [53], considered stereopsis and texture. The data were modeled using a weighted linear combination. Weights were found to be dependent on the viewing distance: in case of short viewing distance, the binocular depth was found to be dominant and only a small weight was given to the texture. In case of a father viewing distance, the weight of texture gradient was found to be greater. This was explained by the fact that stereopsis had smaller reliability with higher viewing distance [25].
Buckley and Frisby [54], also studied binocular disparities and texture, but also analyzed the interaction with “outline cues”. The term “outline cue” was defined by Clarke et al [75] and describes the following property: “if a rectangular is drawn in the plane of the display, rotating it along the vertical axis makes it appear as a trapeze”. This is related to the linear perspective depth cue. In this study, a truncated textured cylinder was considered. These cylinders had different amplitude of depth and two distinct orientations: vertical, or horizontal. Results have shown that the orientation has an impact on how the depth is perceived. In case of a horizontal cylinder, binocular cues dominate the overall depth perception. In case of the vertical cylinder, for low depth amplitude (3-6 cm) outline and texture dominated the perceived depth, but not for high amplitude (9 cm). In the latter case, binocular cues were again the dominant cue.
Landy et al [55] considered kinetic depth and texture integration. A linear model was used with variable weights. These weights were adjusted based on the reliability of the depth cues. The reliability of the depth cues was determined by previous test results and was found to be dependent on the scenes considered.
Defocus blur and binocular disparity were studied by Wang et al [57]. A paired comparison experiment was conducted, and test participants had to report which of the two stimuli, composed of two natural images, provided the larger depth interval between the sharp plane and the blurred plane. In this study, it was found that depth perception was not affected by the viewing distance between the observer and the blurred plane, but was affected by the distance between the sharp plane and the blurred plane, and then by the disparity gradient. One type of Gaussian blur was considered. The blur, provided an increase of perceived depth, but not enough data are available to study the combination of blur and binocular disparity.
Held et al. [58], considered different combinations of binocular disparities and defocus blur to determine depth discrimination thresholds between two distinct objects. The approach is similar to Cutting and Vishton [25] and determines the perceived depth JNDs, but provides additional results by analyzing the interaction between depth cues. Results show that the contribution of each cue was dependent on the viewing distance, and when the viewing distance is small, binocular depth defines the depth discrimination threshold. On average, in the condition of the experiment, when the viewing distance is higher than 32 cm, the defocus blur defines the depth discrimination threshold. This contradicts the results provided by Wang et al [57] where the contribution of blur was found to be independent of the viewing

distance. An explanation is probably due to the different viewing distances which were considered. Indeed, in [57], the viewing distance was less than 32 cm in 147 conditions out of 155; therefore, according to [58] binocular disparity was dominant and explains that in their results [57] the disparity gradient was found to be the only factor.

A popular method for linear depth cues pooling is the use of the *Maximum-Likelihood Estimate (MLE)*. The motivation behind this approach is to take into account the reliability of each depth cue into the pooling, as already mentioned in the case of Landy et al's work [55]. The method uses the following rule: $S_i^*$ is an estimate of the depth $S$ due to the depth cue $f_i$. If the estimation error of $S_i^*$ is a Gaussian noise with a variance $\sigma_i$, then the combination of the $N$ depth cues can be performed by the equation 2.3. This equation gives a higher weight to the reliable depth cues, the ones having a Gaussian noise with a low variance, than to the unreliable depth cues: the ones having a Gaussian noise with a high variance.

$$S^* = \sum_{i=1}^{N} w_i \cdot S_i^*, \quad where \quad w_i = \frac{1/\sigma_i}{\sum_{i=1}^{N} 1/\sigma_i} \tag{2.3}$$

Ernst and Banks [59, 76] studied the combination of binocular and haptic cues and found that MLE could very well predict the integration of the two considered depth cues by the visual system. Hillis et al [60] extended this previous study by showing the validity of the MLE approach for the combination of binoculars and texture gradient cues. This was successfully extended further by Lovell et al [61] to the combination of shading and binocular cues.

#### 2.3.2.2 Strong fusion

In the following, successful applications of strong fusion models to depth prediction are summarized. Evidence of strong fusion can also be found in the literature showing the different types of interaction as listed in subsection 2.2.2.4.

Cue promotion

Rogers and Collett [56] studied the relationship between motion parallax, binocular disparity and the overall depth. Based on an experiment involving motion parallax and zero binocular disparities, it was found that perceived depth revealed by subjective data correspond to half of the actual depth which was defined by design. Different combinations of monocular and binocular depth cues show that parallax affects the perceived depth only when the disparity gradient was small. Similarly, in Johnston et al [53], binocular depth was found to be dominant. A linear model was proposed, and the overall depth is defined as the summation of the binocular depth plus a weighted contribution of the motion parallax. The weight of the motion parallax is defined as inversely proportional to the binocular depth. This reveals the interaction of the type “cooperation” as defined in subsection 2.2.2.4.

Based on the results presented by Bruno and Cutting [52] showing a linear combination of independent depth cues, Massaro [62] suggested another approach called “fuzzy logical model of perception” (FLMP). The motivation behind this new model is the lack of evidence for additivity and individual processing of each depth cue in their results, and therefore, of the weak fusion approach. The strong fusion model proposed is a multiplicative one with a normalization of the different depth cues. The normalization is defined such that the most reliable depth cue has the strongest effect on the overall depth.

Landy et al [49] defined an intermediate model between weak and strong fusion and called it “Modified Weak Fusion (MWF)”. The motivation behind this algorithm is that weak fusion already performs well with modeling subjective data and have a low complexity. The MWF is linear and considers the independence of the different depth cue. Due to this assumption, it may be categorized among the weak fusion models; however, one type of strong interaction is allowed: the cue promotion, which makes it part of the strong fusion schemes.

Cue vetoing

Several studies also report examples of “cue vetoing”. Bülthoff and Mallot [20] analyzed the interactions between binocular depth and shadings. First, regarding binocular depth cues, it was found that the presence of edges is important, and depth perception is considerably reduced in case of smooth disparate images. The contribution of binocular depth to the overall depth is much stronger than shading, but shading was still found to be affecting the depth perception. It was observed that edge-based stereo overrides both shape-from-shading and shape-from-disparate-shading. A conflict between the information from shading and disparity does not result in a veto of the depth information from shading, but results in a reduced depth perception of approximately 25%. Olge [63] demonstrated an example of cue veto using vertical magnifier lenses in front of one eye and altered vertical but not horizontal disparities. This resulted in making the plane surface under study appear as slanted. The increase of the magnification is monotonous with the increase of the slant. However, with a too high conflict between vertical and horizontal disparities, the slant appears to be null again, and therefore it is, an example of the cue vetoing.

Cue disambiguation

An example of “cue disambiguation” can be found in the work of Braunstein et al [64] who have addressed the kinetic depth and occlusion. In this study, they found that occlusion can be used to disambiguate the perceived motion information and can be used to disambiguate the overall 3D-geometry of the surface under study. Andelson [77], reports similar properties explaining that motion parallax can be ambiguous; indeed, the change of speed can be induced by two possible reasons: a change of shape or a change velocity. As a result, there is then an infinity of potential position in depth due to different combination of motion and shape. Besides, motion parallax itself cannot provide information about depth, and other cues such as binocular depth cues are needed. This kind of analysis can also be extended to the relative size depth cue, where it is not always easy to differentiate a change of size due to the size of the objects themselves and their position in depth. Another proof of disambiguation between secularity and shading can be found in the work of Blake and Bülthoff [65].

### *2.3.3 Modeling*

Previous sections have dealt with aspects of depth cue integration. In the following, concrete depth models are presented. To perform the depth cues pooling taking into account all the different possible kinds of interaction, Bayesian models have become the reference for sensory input pooling [67]. Yuille and Bülthoff [66] have described the overall framework and use it on two separate examples, one combining texture gradients and shape from shading, and the other one on binocular depth cues and motion parallax pooling. Due to the importance of these types of models, there are described here, as well as how they work and how they can be applied in practice. The notation provided by Yuille in [66] is used. The basic Bayesian formula is given by Equation 2.4.

$$P(S|I) = \frac{P(I|S)}{P(I)} \tag{2.4}$$

$S$ is the characteristics of a scene such as how different objects are perceived to have distinct position in depth and/or shape. $I$ is the retinal image. $P(I|S)$ is the *likelihood function* for a scene, and is the probability of seeing the image $I$ in case of the characteristics $S$. $P(S)$ is the *prior distribution*, which is the probability to see the property $S$ in the world in general. $P(I)$ can be seen as a normalization constant. $P(S|I)$ is the *posterior distribution* and describes the probability of the depth characteristic $S$ to be seen in the retinal image $I$. In order to get the depth estimate $S^*(I)$, it is needed to find the value of $S$ which will maximize the *posterior distribution*, $P(S|I)$, as express by Equation 2.5. This value is called the *maximum a posteriori (MAP)*.

$$S^* = arg\ max_S P(S|I) \tag{2.5}$$

Based on this formalization, the process of estimating depth from two depth cues $f$ and $g$ can be performed in a weak or a strong manner. In the case of the weak fusion model, the depth estimates $S_1^*$ and $S_2^*$ are determined. First, $S_1^* = arg\ max_S P(S|f)$ and $S_2^* = arg\ max_S P(S|g)$. Then, a weak fusion model combines $S_1^*$ and $S_2^*$ using, for example, using a weighted mean.

The approach relevant to this section is the possibility to perform a strong fusion: the depth estimated from two cues, $S_{1,2}^*$, is determined as defined in Equation (2.6) and is based on the analysis of both cues in a same module, and not two separate modules as presented previously.

$$S^* = arg\ max_S P(S|f,g) \tag{2.6}$$

Equation 2.7 is the application of the two-depth-cue case to equation 2.4.

$$P(S|f,g) = \frac{P(f,g|S)P(S)}{P(f,g)} \tag{2.7}$$

An intermediate step between weak and strong fusion is to consider $P(f,g|S) = P(f|S) \cdot P(g|S)$. This could be possible if the two prior, e.g. the probability to see a certain amount of depth $S_1$ and $S_2$ respectively from a cue $f$ and $g$ are the same. This is a halfway step because the two cues are decoupled and do not follow the definition of a strong fusion algorithm. However, this is not a weak fusion either since the depth estimate $S_{1,2}^*$ is not resulting from a linear combination of $S_1^*$ and $S_2^*$. This result is given in Equation 2.8, which was employed by Girshick and Banks [67] for combining binocular and the texture gradient depth cues. A simplified form of such a model can also be found in Nakayama and Shimojo [50] where prior probabilities, $P(S)$, are neglected and the overall probability of the scene is based on the depth cue which shows the highest probability to explain the scene.

$$P(S|f,g) \propto P(f|S) \cdot P(g|S) \cdot P(S) \tag{2.8}$$

The computation of $P(f|S)$ can be achieved by knowing the properties of the depth cues under study. In the example of Yuille and Bülthoff [66], the probability of light and shades for a retinal image $I$, can be computed based on the imaging model in Equation 2.9.

$$I = \mathbf{s} \cdot \mathbf{n} + \mathbf{N} \tag{2.9}$$

Here $\mathbf{s}$ is the light source, $\mathbf{n}$ is the normal to the surface, and $\mathbf{N}$ is a Gaussian noise. The likelihood function is $P(I|S) = (1/Z)e^{-(1/2\sigma^2)(I-\mathbf{s}\cdot\mathbf{n})^2}$ where $Z$ is a normalization factor, and $\sigma^2$ is the variance of the Gaussian noise [49]. Such model needs to be defined for each depth cue and enables to find the most appropriate solution. However, as mentioned by Yuille and Bülthoff [66], the hypothesis made for defining the likelihood function has a strong influence on the results and should carefully be taken, considering that hypotheses for different depth cues can be contradictory as in case of shading and texture gradient.

Another important aspect is the use of Bayes' decision model in the optimization process for finding the best solution. This is done through the definition of a loss function. This function defines the cost of making a prediction error for each depth cue. Let $L(S,d)$ be the cost of estimating the value $d$ instead of $S$. The risk function is defined in Equation (2.10).

$$R(d) = \int L(S,d)P(S|I)dS \tag{2.10}$$

Finding the best solution results in finding the value $d^*$ which minimizes the risk. The advantage of such practice is to be able to provide a different weighting for each depth cue. This enables taking into account the reliability of one depth cue in the process of finding the best solution by defining the loss function per depth cues. Such application can be found in Girshick and Banks [67], where they found that texture gradient was not as reliable as binocular depth cues and the contribution of the two cues was depending on how they agree: for small disagreement, both cues are considered, and for strong disagreement texture gradient is discarded. These results can be explained by the Bayes' decision model with an appropriate risk function.

### *2.3.4 Conclusion*

In this section, a comprehensible review of the work performed on depth perception in psychophysic was provided, addressing the questions: What are the different depth cues which will be further studied along the thesis, how they relate to each other, what are the different models for predicting depth perception. The next section will address briefly a second important aspect: the visual comfort.

## 2.4 Visual comfort

Visual comfort is one of the biggest issues with regard to 3D Quality of Experience, as stated in the introduction. *Visual discomfort* is usually related to *visual fatigue*. *Visual fatigue* is related to “the decrease of performance of the human visual system” and the *visual discomfort* is “the subjective counter-part of visual fatigue” [15]. Urvoy defined the visual discomfort as a factor that can be evaluated subjectively, while visual fatigue is rather a symptom that can be evaluated through objective measurement in clinics by doctors [78]. In both definitions, the visual fatigue relates to the long-term effect of visual discomfort.

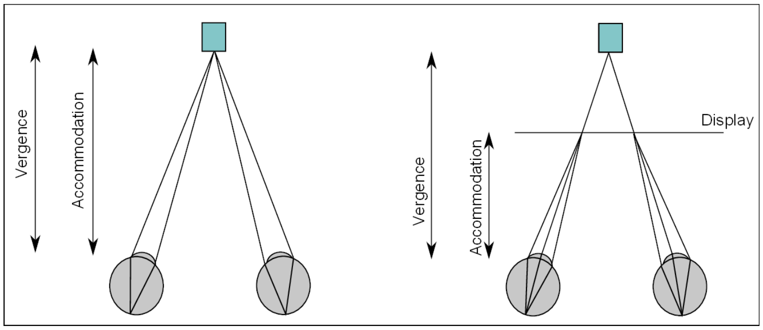


Figure 2.28: Vergence accommodation conflict.

A large amount of research has been carried out to evaluate the effect of the two last-mentioned factors on the visual fatigue. The typical theorized reason of visual comfort is the vergence accommodation conflict [79, 80]. Figure 2.28 illustrates this problem. It relates to the phenomenon of eye convergence and accommodation which is usually updated in a synchronized manner. However, in the case of 3DTV, the focus has to be kept on the screen and the vergence follows the 3D rendered object in depth. This results in a disagreement between accommodation and vergence which is unnatural for the human visual system. However, there are contradicting results with regard to this explanation, and it has to be noted that studies have found that the accommodation does not always remain focused on the screen but can also move to the 3D object [81]. But it is not clear if the observed shift of accommodation from the display plane is due to the change of vergence or natural underaccommodation which occurs in case of near objects [15, 82].

Moreover, the accommodation does not always need to be updated: there is an area called the *depth of focus* (DOF) which corresponds to an area where vergence could change while keeping the object sharp without changes of accommodation. The DOF ranges from 0.04 to 3.50 diopter, and has typical values from 0.2 to 0.5 diopter [15]. Within this area, it is then possible to have an update of the vergence while the accommodation stays the same without being perceived as unnatural to the human visual system. Based on this area a comfort zone was defined, this area states that retinal disparities should be less than 1° [83].

Alternatively, a zone of comfort was proposed by Percival and is called the *Percival area*, and can be seen as an alternative to the 1° rule [15]. It is based on the middle third of the amount of binocular vergence with almost no change in accommodation [15, 84]. However, there is lack of agreement on how to define this area: some studies used break points whereas others used blur points. The representation of the different zones depending on how they were determined are illustrated in Figure 2.29.

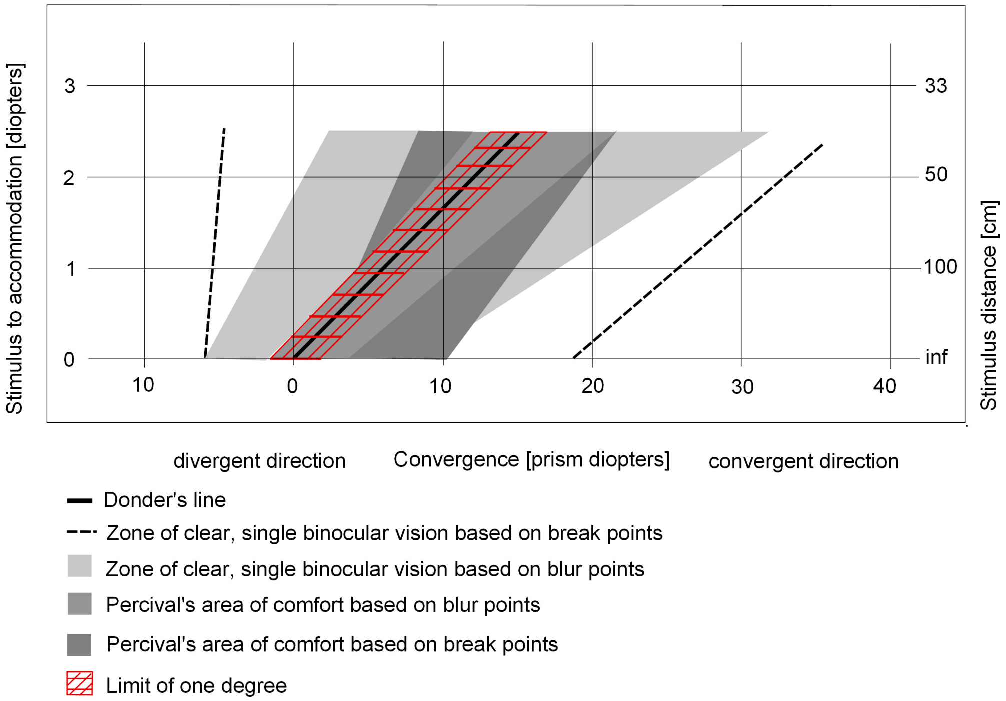


Figure 2.29: Different comfortable viewing zones. This describes the different zones describing where single binocular vision and comfort can be achieved as a function of retinal disparities and distance to the stimulus. These areas depends on how the evaluation was performed (break points or blurred points). (Figure from [15])

Within the comfortable viewing zone, visual discomfort might still occur in case of too much variation of disparities [85]. This was found to be the case for scenes having large amounts of disparity and motion. Moreover it was observed that discrete changes of motion in the depth direction in stereoscopic sequences results in a decrease of the accommodation response and a significant decrease of visual comfort [86], as reported in [15]. In the particular case of the alternation between positive and negative disparity was also found to have a high effect on visual comfort [87]. As presented in this section, there are many different sources of visual discomfort. To limit discomfort issues, different recommendations, were provided in the literature. In [83] an upper limit of 70 arcmin is recommended for retinal disparity, in ITU-R Recommendation BT.1438 [88] 0.3 Diopter corresponding to 60 arcmin is suggested. This value was also supported in [15] since until this limit sharp binocular binocular vision is preserved, and blurred images are expected to be a first step before seeing double and suffering of discomfort. Another rule of 0.2 Diopter was also proposed in [86, 89] considering that discomfort is clearly perceivable outside of 60 arcmin (0.3 Diopter) area. In parallel, in professional shooting, the 1/30$^{th}$ rule of thumb for 3D production is usually used and states that the inter-camera distance should be 1/30$^{th}$ of the distance from the camera to the first foreground object [90]. However, this method is empirical since it only roughly considers cameras and display configuration.

## 2.5 Technical implementation

After presenting a review of depth perception, this section will address how 3D contents are captured and rendered on a stereoscopic display. This will show the limit of 3D rendering technologies and will be put into relation with the perceptual factors presented in the last section. This section will show that both capture and rendering are closely related and should both be taken into account to ensure an appropriate depth effect.

### *2.5.1 Capture*

When considering the case of shooting a stereoscopic sequence (which can be extended to N views), two choices for the setup of the cameras are possible to achieve high depth quality: adjust the optical axes to be parallel, or converging to the object under focus (see Figure 2.30). Parallel camera axes during shooting require setting the convergence during post-production, which can be time consuming. Having the camera axes converge during shooting requires time during production, but less time in post-production. However it can create keystoning [91] issues which need to be corrected in post-production. Both approaches are equally valid and used in the context of stereoscopic 3D production.

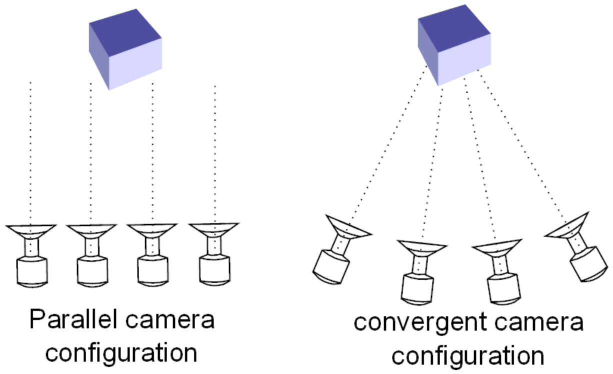

Parallel camera configuration
convergent camera configuration

Figure 2.30: Different type of camera's configuration

In the following, the case of converging cameras is considered, however similar properties can be found in the case of a parallel camera setup. Figure 2.31 depicts a configuration where two cameras record one object. Following optical geometry properties, the image of the object is projected on each sensor of each camera at a different position. The difference in the position of the projected image on each sensor depends on different factors: the inter-camera distance, the position of the object relative to the camera, and the focal length. Additionally, to convert the distance in meters of the projected images on the sensor to difference in pixels, two other parameters need to be considered: the sensor size and resolution. In this figure, one particular issue can be observed: the discretization of the depth into the limited number of pixels of the sensor. Indeed, depending on the setup of the two cameras and the focal length of the lenses, the position in depth result in a value of difference of position of the projected image on the cameras' sensor. However, since the camera sensor has a limited resolution, the continuous image projected on the sensor by the lenses will result in a discretization of the image, and thus a discretization of the difference of position in depth. Therefore, due to the limited resolution, there will always be a point where a difference of depth cannot be recorded because the difference of position of the projected image on the sensor is too small.

In addition, a major issue about the quality of the depth is the choice of camera settings with respect to the display settings. Due to optical properties, the choice of a specific focal length of a camera to shoot an object at a specific

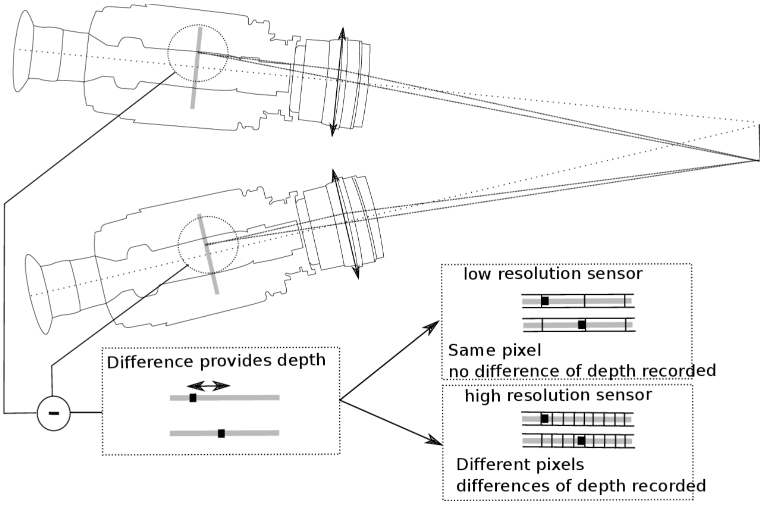

low resolution sensor
Same pixel
no difference of depth recorded
high resolution sensor
Different pixels
differences of depth recorded
Difference provides depth
-

Figure 2.31: Conversions of depth recorded by the cameras to pixel differences and link with sensor resolution

distance has an effect on the lines inside the picture and affects its geometry. This phenomenon is illustrated in Figure 2.32, where the same scene is recorded with different focal lengths. The process of rendering a 3D image is the result of capturing and rendering the scene. Hence it is the result of three successive conversions: World in the camera's space $\Longrightarrow$ pixels on the camera sensor $\Longrightarrow$ pixels on the display $\Longrightarrow$ presentation in 3D in the user's space.

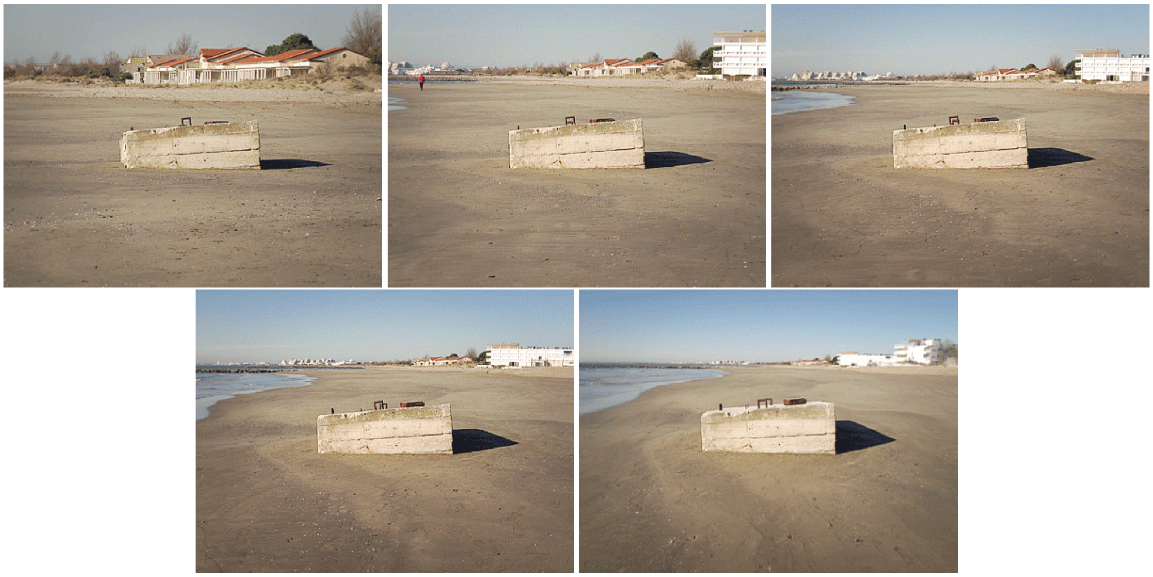


Figure 2.32: Relation between object geometry and focal length (images from Micaël Reynaud)

The relation between individual capture settings and their result on a rendered depth on a 3D display was studied by Woods et al [91], and Figure 2.33 depicts how changes between the different parameters keeping the other parameters constant affect the geometry of the 3D rendering. What has been captured by the camera is a rectilinear grid, which is significantly distorted by the end-to-end system.

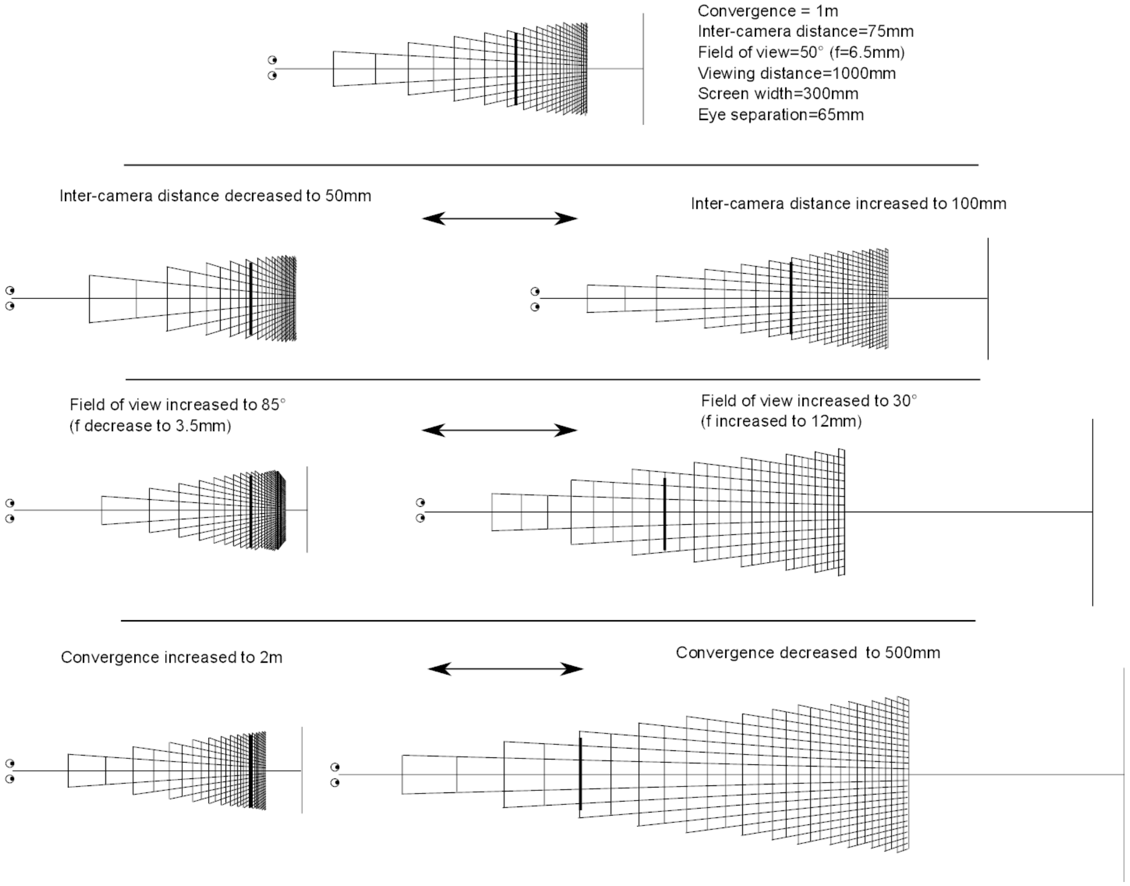

Convergence = 1m
Inter-camera distance=75mm
Field of view=50° (f=6.5mm)
Viewing distance=1000mm
Screen width=300mm
Eye separation=65mm
Inter-camera distance decreased to 50mm
Inter-camera distance increased to 100mm
Field of view increased to 85°
(f decrease to 3.5mm)
Field of view increased to 30°
(f increased to 12mm)
Convergence increased to 2m
Convergence decreased to 500mm

Figure 2.33: Relation between camera settings and 3D rendering (Figure redrawn from [91])

Figure 2.34 depicts how the depth is rendered as a function of the distance from the camera when the parameters regarding the rendering is fixed. Ideally, a linear relationship between the depth of the camera space and the depth in the display space should exist. Due to the previously mentioned limitations, this is not the case, and it can be seen that the area where the distortion is minimal appears to be only limited to a reduced area in the camera space. The relevant part of the movie or image should then be located in this particular area to ensure a good depth quality rendering.

As described in [91, 24], the depth in the visualization space can be expressed by Equation (2.11). With:

- V: the viewing distance, e.g.. the position of the viewer in front of the display
- B: the interpupillary distance
- z: the position of an object in depth in the camera space
- M: the magnification factor, e.g.. the ratio between the size of the camera sensor and the display size
- f: the focal length of the camera
- $d_{cov}$: the distance from the camera to the convergence point, e.g. the zero depth plane
- Z: Depth in the visualization space

$$Z = \frac{V \cdot B \cdot z}{B \cdot z + M \cdot f \cdot b(1 - \frac{z}{d_{cov}})} \tag{2.11}$$

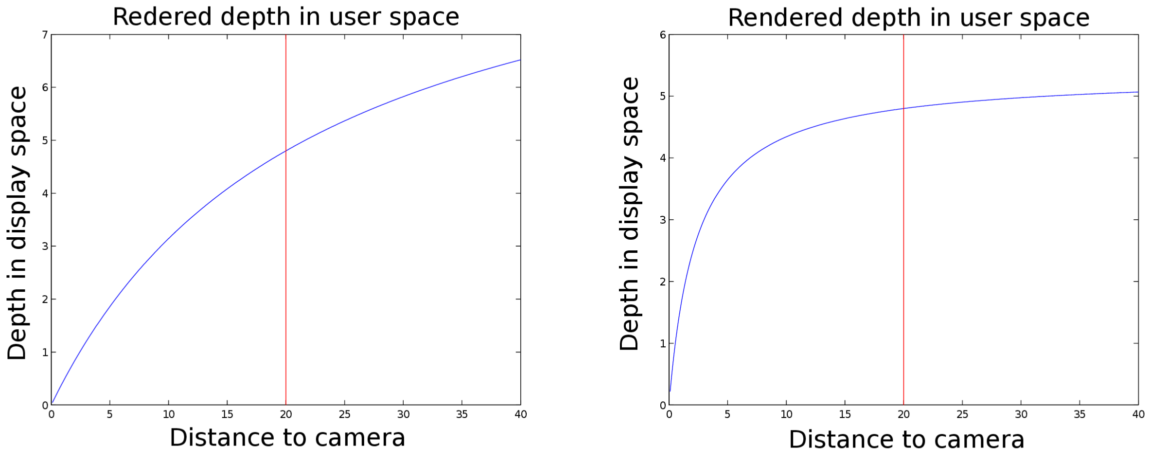


Figure 2.34: Relationship between camera and display space as a function of camera focal length

### 2.5.2 Rendering

The rendering capabilities also are of high importance to ensure a high-quality depth rendering. In the previous section, Figure 2.34 was used to illustrate the effect of camera settings on the quality of the depth rendering. However, one important aspect was only briefly mentioned: these distortions also depend on the resolution of the display, the viewing distance and the inter-pupilar distance. These parameters and their effect on the 3D rendering were studied by Woods [91], and the effect of these different settings on the depth rendering is illustrated in Figure 2.35. Moreover, displays have a limited resolution which results in a quantization of the 3D rendered depth in 3D pixels called voxels. Figure 2.36 illustrates this quantization of the depth on a display.

### 2.5.3 Transmission

In between capture and rendering, one important aspect is the transmission of the 3D material. This includes encoding and transmitting the bitstream, over an IP network. Due to the amount of data contained in a 3D-video and the limited bandwidth of networks it is necessary to encode the 3D videos. Different alternatives are available depending on the needs in terms of the number of views and backward compatibility. There are different current encoding strategies which will be explained in more details in the following. These are:

- Frame packing and a 2D-video encoding
- Simulcast
- Multi view coding (MVC)
- Depth-based video encoding

The first approach, frame packing and a 2D-video encoding was chosen for early-days IPTV broadcast since it ensures compatibility with the legacy transmission chain. In this particular case, the two stereoscopic views are packed into a single frame. Different ways to perform the packing are possible, these include having two frames one over the other, next to each other, or interlaced (see Figure 2.37). The frame packing can also be done jointly with a downsampling of the video resolution in the direction of the packing in order to keep the same frame size as one of the individual views, but at a cost of images sharpness. However, this will enable to bring the 3D video to the end users without adding additional costs in terms of bandwidth. Once packed, the videos are then encoded using a traditional 2D video encoder such as H.264.

A second approach is called "simulcast", it consists of encoding both stereoscopic views separately using a 2D video

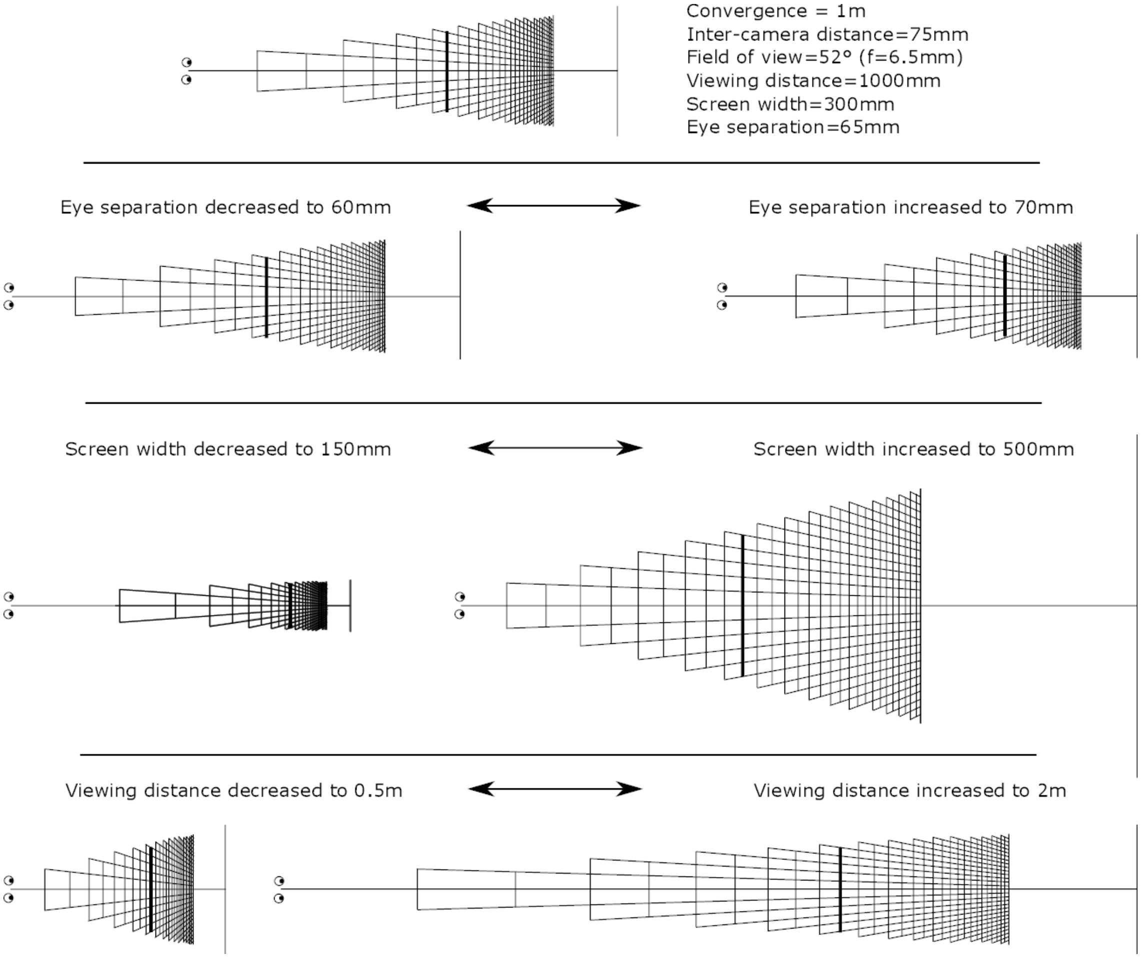

Convergence = 1m
Inter-camera distance=75mm
Field of view=52° (f=6.5mm)
Viewing distance=1000mm
Screen width=300mm
Eye separation=65mm
Eye separation decreased to 60mm
Eye separation increased to 70mm
Screen width decreased to 150mm
Screen width increased to 500mm
Viewing distance decreased to 0.5m
Viewing distance increased to 2m

Figure 2.35: Relation between visualization settings and 3D rendering (Figure redrawn from [91])

encoder. The two bitstreams are transmitted separately resulting in a doubling of the required bitrate.
The third approach was specifically designed for 3D videos and uses the redundancy between the views. The multi-

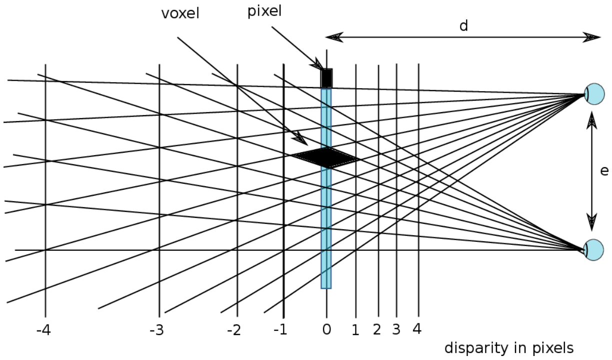


Figure 2.36: Depth rendering and 3D voxels (Figure redrawn from [92]). *e* is the inter-pupillary distance, *d* is the viewing distance.

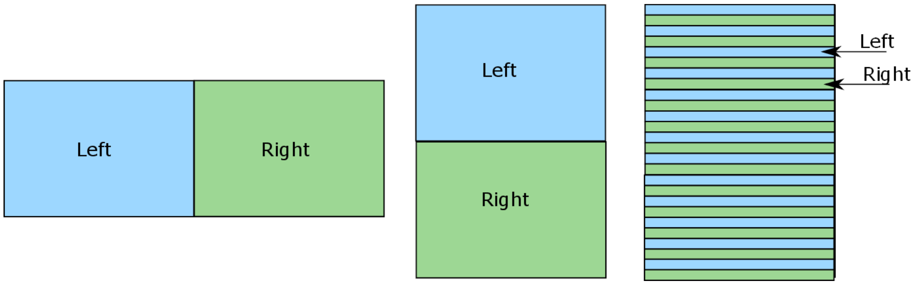


Figure 2.37: Different methods for packing two views into a single frame.

view coding is composed of one stream for one view, and the other view is encoded relative to the first view. During the encoding, the encoder can choose to encode the P- and B- frames relative to the previous I- or P- frames or use a frame from the other view as reference (Figure 2.38). With such an approach, the compatibility with legacy transmission chains is maintained. The information from the second stereoscopic view is dropped by the decoders which do not handle the MVC bitstream, thus enabling to decode the sequence as 2D video. Such approach usually saves 20% of the overall bitrate compared to simulcast. MVC is currently used in the industry in 3D Blu-rays.

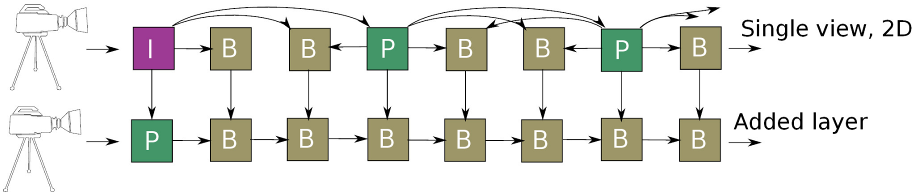


Figure 2.38: Relationship between frames in a MVC bitstream.

A fourth alternative are depth-based video coding strategies. These methods enable much higher coding performance than the other presented alternatives. It requires one view, the "texture", a depth map, and information about the camera setup. This information is transmitted to the client which "synthesizes" the missing view. There are different sub-categories within this category: video plus depth (V+D) which uses one view and a depth map to synthesize a missing stereoscopic view (Figure 2.39). However, since the newly reconstructed frame has a different point of view than the existing frame, it contains areas which were occluded and need to be filled. Filling these holes can be challenging, that is why alternative depth-based coding methods have been proposed. The multi-view plus depth (MVD) was then proposed. It has different texture maps which come from different cameras having different points of view. These textures, in addition to one or more depth map, can then be used to solve the problem of filling holes ("inpainting"). Still with the limitations of inpainting performance, an alternate format is the layered depth video (LDV), which similarly to MVD, uses different texture maps. Redundancy is reduced by specifically containing the background, which was hidden by the foreground object (Figure 2.40).

## 2.6 Conclusion

This chapter presented the background of this thesis. It showed different alternatives from the literature on how 3D quality of experience can be evaluated based on different evaluation concepts. The relation between these concepts were shown across different models of QoE. In the work conducted on 3D QoE, a number of studies have been performed on the evaluation of 3D image quality, visual discomfort, and depth perception. But the last one, depth from

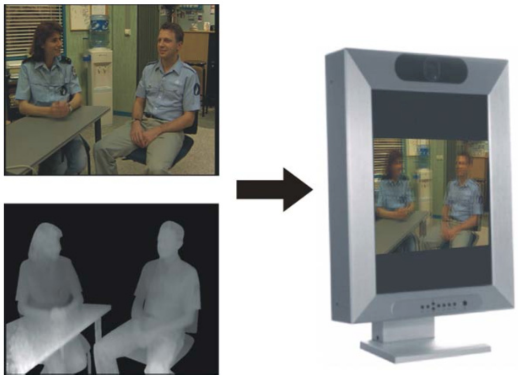


Figure 2.39: Video plus depth (V+D) coding. Figure copied from [93]

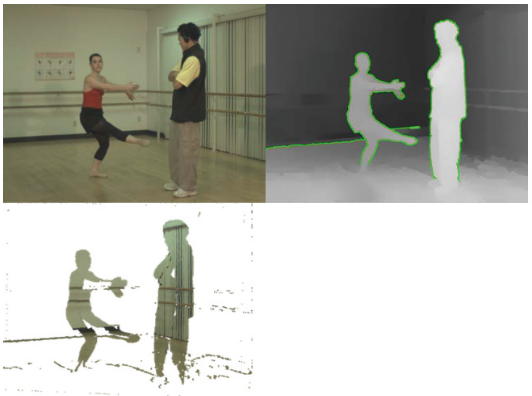


Figure 2.40: layered depth video (LDV) coding. Figure copied from [93]

different depth cues and its relation with QoE, received less attention than what was presented from the literature in this chapter. This chapter has presented an extensive review of depth perception based on binocular and also monocular cues. The monocular cues were until now less considered in 3D QoE studies, which is why work from this area has been included in more detail in this thesis. Since our tests, have been performed using 3D displays, it was also necessary to present the effect of technical factors such as capture and rendering settings on the perception of depth.

The next three chapters of this thesis will describe the contribution of this thesis: to study the relationship between QoE and depth, from different perspectives. Depth quantity will be studied specifically, and the question of how to evaluate the contribution of different depth cues both using subjective tests and algorithms for depth cue prediction.

# Chapter 3
# Evaluating 3D added value

Chapter 2.1 described the different notions of Quality of Experience (QoE), perceived depth, and image quality. This chapter targets the evaluation of the added value of 3D video as compared to 2D in case of different transmission scenarios. The goal is to show how the depth, the image quality and QoE relate to each other. As explained in the previous chapter on the state of the art, measuring the differences between 2D and 3D in terms of QoE can be challenging and depends on the context. Hence, in this chapter it will be described how different evaluation methods have been considered in this thesis, and their limitations are analyzed based on the obtained results. Finally, to enable revealing the differences in terms of QoE between 2D and 3D, another evaluation paradigm will be used. Using this other evaluation methods, it will be possible to relate the added value of 3D to the depth effect revealing the need to characterize the properties of the source material. The structure of this chapter is illustrated in Figure 3.1, as well as their results. And Table 3.1 provides a list of the experiments described in this chapter.

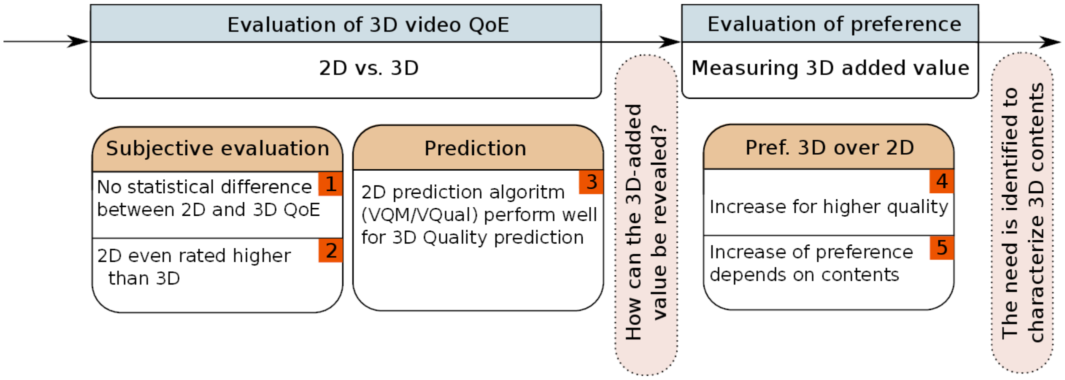

Evaluation of 3D video QoE
2D vs. 3D
Subjective evaluation
No statistical difference between 2D and 3D QoE
1
2D even rated higher than 3D
2
Prediction
2D prediction algoritm (VQM/VQual) perform well for 3D Quality prediction
3
How can the 3D-added value be revealed?
Evaluation of preference
Measuring 3D added value
Pref. 3D over 2D
4
Increase for higher quality
Increase of preference depends on contents
5
The need is identified to characterize 3D contents

Figure 3.1: Structure of the studies described in the chapter.

| Section | What is evaluated | Type of degradations | Methodology | Published in |
|---|---|---|---|---|
| 3.1.1 | 3D QoE and visual comfort | Video encoded with different coding schemes, bitrate, and transmission impairments | ACR | [3] |
| 3.1.4 | 3D QoE and 2D video quality prediction model accuracy on predicting 3D quality | Video encoded at different bitrate | SAMVIQ | [94] |
| 3.2 | 3D QoE and preference of 3D over 2D | 2D and 3D videos encoded at different bitrate | Pairwise comparison | [95] |

Table 3.1: List of experiments conducted to address the evaluation of 3D QoE.

## 3.1 Differences and similarities between 2D and 3D QoE for streamed videos

The two first studies which will be presented target the case of an IPTV service. The main goal is to evaluate the QoE of a user faced with the proposed video streaming service. In such a case, different aspects are to be considered: the contents, how contents are encoded and how transmission errors and their concealment affect the QoE. To this aim, two experiments were conducted with different contents, coding algorithms and under different network conditions. In every case, test participants were involved and had to report QoE scores for each stimulus providing insight into their experience of the service.

### *3.1.1 Subjective evaluation of 2D and 3D QoE*

In this first experiment the differences between QoE scores between 2D and 3D videos were compared in an experiment involving test participants. The research questions were the following:

**Listing 3.1: Research questions**

```
1. Compare 2D and 3D QoE on the reference video, where no degradation was applied.
2. Compare how 2D and 3D video sequences are rated when encoded in the same manner, and
   attempt to measure the added value of 3D compared to 2D video sequences.
3. Compare different coding schemes: Simulcast, MVC, and Side by Side representation
   with H.264 encoding.
4. Compare 2D and 3D QoE in case of an error-prone transmission chain impacted by packet
    losses. Such errors resulting in ``slicing distortion''.
5. Study visual comfort and its link with QoE ratings.
```

#### 3.1.1.1 Source material

Seven contents have been used as source material (SRCs). All of them were 10s long full-HD progressive sources of 25 frames/second (1080p25). The contents have different spatial, temporal and depth characteristics as summarized in Table 3.2.

| Content Name | Description |
|---|---|
| Horse | Horse standing in a field, scene change, car approaching, scene change, the horse starts to walk.<br>This content has complex texture and a slow pan motion. |
| Car Race Prep. | Preparation of a race; several scene changes, colorful, high spatial complexity, slow motion. |
| Car Race | Scene with cars racing; several scene changes, high motion and large depth range. |
| Piano | Man playing the piano; slow pan motion, low spatial complexity. |
| Ski | Skier skiing; low on texture, high motion, large depth range. |
| SkullRock | 3D generated sequence, low spatial complexity, low motion, and high depth range. |
| Boxe | Two men boxing. There is only the boxers' movement. |

Table 3.2: Source sequences characteristics used in the study.

#### 3.1.1.2 Processing of test sequences

To generate all the Processed Video Signals (PVS), several Hypothetical Reference Circuits (HRCs) were considered. These HRCs can be divided into two distinct groups: coding-only and coding under transmission errors. The general

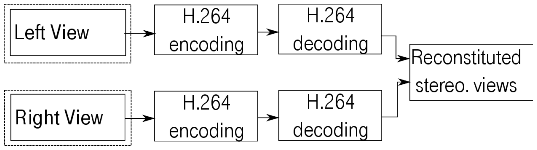


Figure 3.2: Processing chain simulcast.

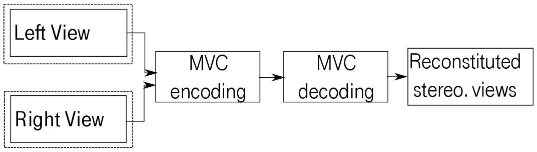


Figure 3.3: Processing chain MVC.

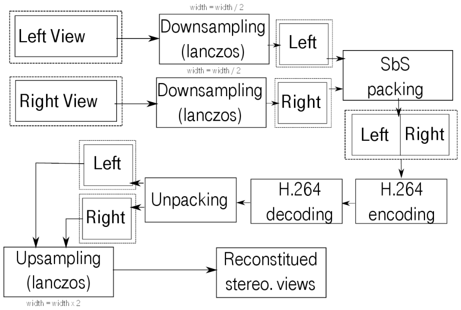


Figure 3.4: Processing chain Side by side.

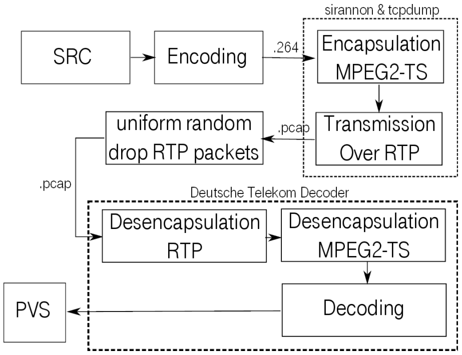


Figure 3.5: Simulation of transmission errors

| HRC | Coding Scheme | QP | Packet loss rate [%] | HRC | Coding Scheme | QP | Packet loss rate [%] |
|---|---|---|---|---|---|---|---|
| 1 | Simulcast | - | - | 13 | MVC | 40 | 0.0 |
| 2 | Simulcast | 26 | 0.0 | 14 | Frame Packing (SbS) | 26 | 0.0 |
| 3 | Simulcast | 26 | 0.4 | 15 | Frame Packing (SbS) | 32 | 0.0 |
| 4 | Simulcast | 26 | 0.9 | 16 | Frame Packing (SbS) | 38 | 0.0 |
| 5 | Simulcast | 32 | 0.0 | 17 | Frame Packing (SbS) | 40 | 0.0 |
| 6 | Simulcast | 38 | 0.0 | 18 | 2D | - | - |
| 7 | Simulcast | 38 | 0.4 | 19 | 2D | 26 | 0.4 |
| 8 | Simulcast | 38 | 0.9 | 20 | 2D | 26 | 0.9 |
| 9 | Simulcast | 40 | 0.0 | 21 | 2D | 38 | 0.0 |
| 10 | MVC | 26 | 0.0 | 22 | 2D | 38 | 0.4 |
| 11 | MVC | 32 | 0.0 | 23 | 2D | 38 | 0.9 |
| 12 | MVC | 38 | 0.0 | | | | |

Table 3.3: Hypothetical Reference Circuits (HRCs)

processing procedure according to the first part of the HRCs is described in Figures 3.2 - 3.4, where encoding is done according to one of three different coding schemes:

1. Simulcast (Figure 3.2 ): The two views are encoded independently using an H.264 encoder (x264 [96])
2. MVC (Figure 3.3 ): The two views are encoded exploiting the redundancy between views, here using JMVC 8.2.
3. Side by Side and H.264 (Figure 3.4 ): The two views are each downscaled and encapsulated in an HD frame, then encoded using H.264 (x264).

For each coding scheme, different values of Quantization Parameter (QP) have been chosen. Defining QP instead of bitrate enables to reach a more constant quality over all SRCs and thus avoids having contents with always low quality (because the maximum bitrate is too low to achieve high quality) or having contents with always high quality (because the selected range of bitrates always leads to high quality). Details on the different HRCs are listed in Table 3.3.

The second part of the HRCs covers different conditions of transmission errors. The process of simulating the transmission errors is depicted in Figure 3.5. Each view of the SRC is encoded independently using an H.264 encoder (x264). The frames were decomposed into 68 slices, which corresponds to one slice per macroblock line. The GOP structure was (M,N) with M=3 and keyframe rate N=1/s. The software "sirannon" [97] was used to encapsulate the bitstream into MPEG2-TS packets, and the resulting TS-packets into RTP packets. The software tcpdump is then used to capture the RTP packets and save them in a packet capture file (".pcap"). The simulation of the lossy channel was performed as follows: we used a random number generator which indicated us the packet number which should be dropped. This random number generator followed a uniform law, and no content-dependent difference between RTP packets were made (e.g. in terms of whether an I-, P- or B-frame was hit by the loss). We only took care that the first I-frame of the PVS was not affected by packet loss. Finally, we used a decoder implemented by Deutsche Telekom Laboratories to decode the video. This decoder (used by ITU-T SG.12 in the context of the P.NAMS and P.NBAMS standardization contests now ITU-T Recommendation P.1201 and P.1202) has the particularity to be able to take pcap-files as input. This decoder implements an intra-error concealment algorithm. The 2D sequences have been realized by presenting the same (left) view to the two eyes.

#### 3.1.1.3 Subjective experiment

For the subjective experiment, the laboratory test environment was set as defined in ITU-R BT.500-12 [98]. A 23" Alienware OptX 3D Full HD Display was used. This display has a native resolution of 1920x1080 pixels and a refresh rate of 120Hz. The display was used in combination with NVidia 3D Vision shutter-glasses. The viewing distance was set to three times the picture height (3H). The maximum value of crossed and uncrossed disparities were checked on every SRC (using a motion estimation-based algorithm to estimate stereo disparities [99]. This will be further detailed in chapter 5) to ensure that the disparity values stay in the comfortable viewing zone. The luminance of the background was set to 50 cd/m$^2$.

The test methodology was Absolute Category Rating with hidden reference (ACR-HR). 21 observers took part in the test, and were asked to rate the general Quality of Experience and the visual discomfort, each on a five grades discrete scale with the typical labels "Excellent", "Good", "Fair", "Poor" and "Bad". It is only after rating the PVS on these two scales that the observers were allowed to watch the next PVS. After screening using the methodology described in the VQEG 3DTV Test Plan, one observer was rejected.

The general procedure of a test was as follows: the test started by a training session composed of seven sequences. This training was designed to illustrate the rating task and to introduce the ranges of contents and of quality. In the main session, the observer could rate the 161 sequences in two sessions, one of 81 and one of 80 PVSs (with a 15min break between the two parts). The whole test (including a vision test and break) took 1.25 h.

#### 3.1.1.4 Relation between coding and 3D QoE

The first objective of our test was to compare the Quality of Experience and consequently bit rate requirements of video encoded with different coding schemes. Here, the Side by Side (SbS) representation currently used for 3D IPTV broadcasting was to be compared with other available algorithms (simulcast and MVC). Figures 3.6 and 3.7 depict the mean quality rating per content and per coding scheme as a function of the logarithm of the bit rate. Also shown are the 95% confidence intervals (CIs). As can be seen from the graphs, CIs are rather high. A MANOVA (Multivariate ANalysis Of VAriance) to explain quality with the fixed factors (QP, coding scheme, contents) was used. No interaction terms were considered, and therefore a remaining of 147 degree of freedom were left. This analysis reveals that there is a significant impact due to coding scheme (F=10.673, $p < 0.01$), a significant impact due to content (F=3.153, $p < 0.01$), and a significant impact due to QP (F=27.33, $p < 0.01$). A non parametric Kruskal-Wallis test applied to explain the MOS values as a function of the coding scheme, shows a significant effect of the coding scheme on the MOS values (chi-squared = 99.032, df = 3, p-value $< 2.2e-16$). Pursuing these results, a post hoc test relating the coding scheme and the MOS values shows that SbS is significantly different than the 2D conditions, and similarly MVC is significantly different than the 2D conditions. However, the 3D conditions were not found statistically different from each others.

From Figure 3.8 we can observed that at a given bitrate level, in most cases SbS provides a higher perceived quality than Simulcast and MVC. However, as analyzed in the previous post hoc analysis, the difference was not found to be significant. We did not observe a strong gain in bit rate for the MVC coding scheme compared to Simulcast. However, an advantage of MVC which is not taken into account in this study is its backward compatibility (e.g.

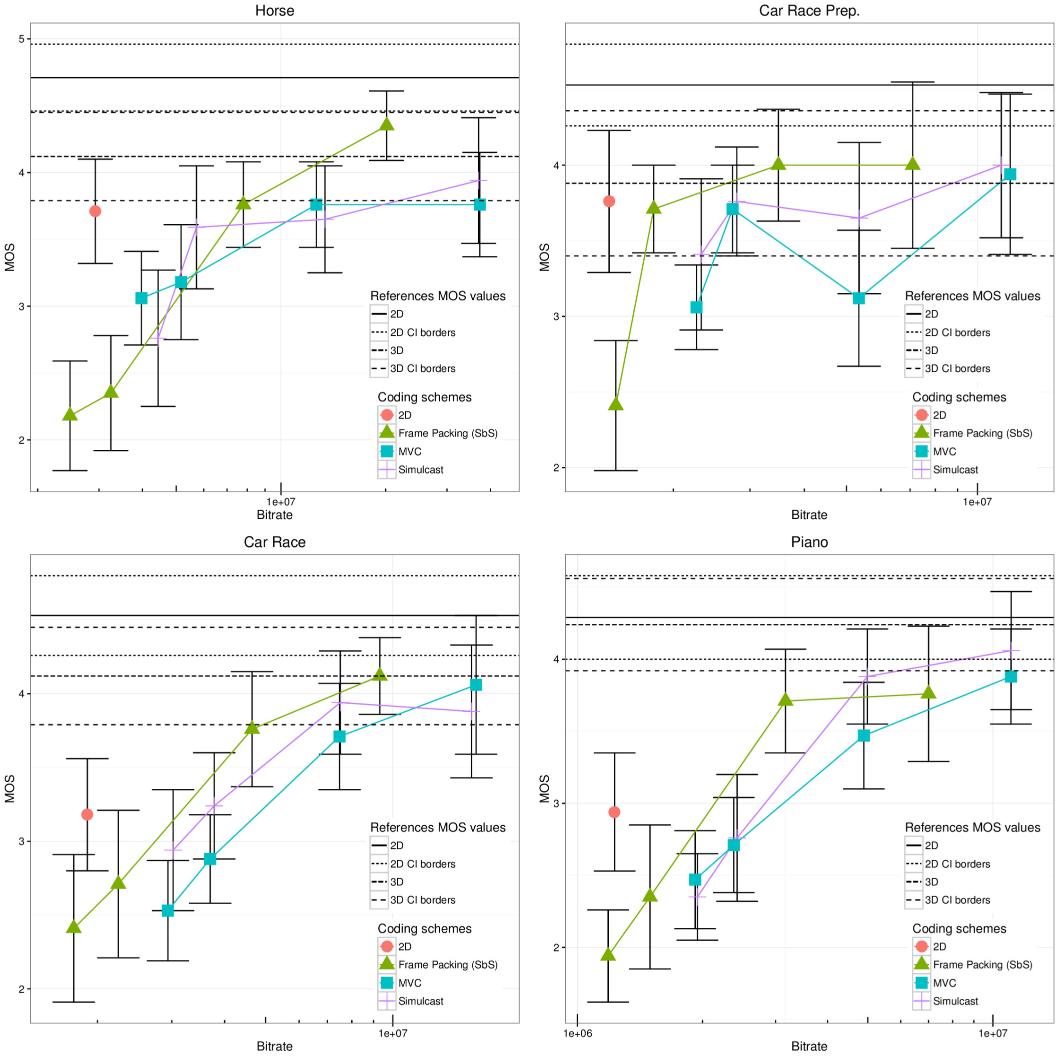


Figure 3.6: Quality per content and per coding scheme (first part of SRCs).

a MVC bitstream can be decoded by a H.264/AVC -compatible decoder simply by dropping the data it does not understand and related to the other views), which is an important feature for 3DTV broadcasting. For evaluating the difference of required bit-rate between methodologies, the approach proposed by Wang et al [100] was used: For every PVS in SbS representation, the value of bitrate required for achieving the same perceptual quality but using Simulcast or MVC is determined. The estimation of equivalent bitrate is done using a linear regression between known values in the log(bitrate) vs. Mean Opinion Score (MOS) space. Then, the ratio of required bitrate for SbS divided by the bitrate required for Simulcast or MVC is calculated. This provides a measure of the relative bitrate gain for each coding scheme. Table 3.4 provides the results in terms of equivalent bitrate. On average, a 50% gain in bitrate can be reached using the SbS representation, without reducing perceptual quality. These results are in accordance with previous tests from the literature [100, 10]. We can also see that MVC did not provide a significant quality improvement in our experiment, and that the results were highly content-dependent. These test results seem to indicate that for a given bitrate the current implementation of 3D HDTV broadcast services achieves a higher quality than the other available standards (using full resolution), when a specific limited value of bit rate is required. We can also observe that in most cases the quality level that can be achieved with SbS is almost as high as the quality achieved with the simulcast

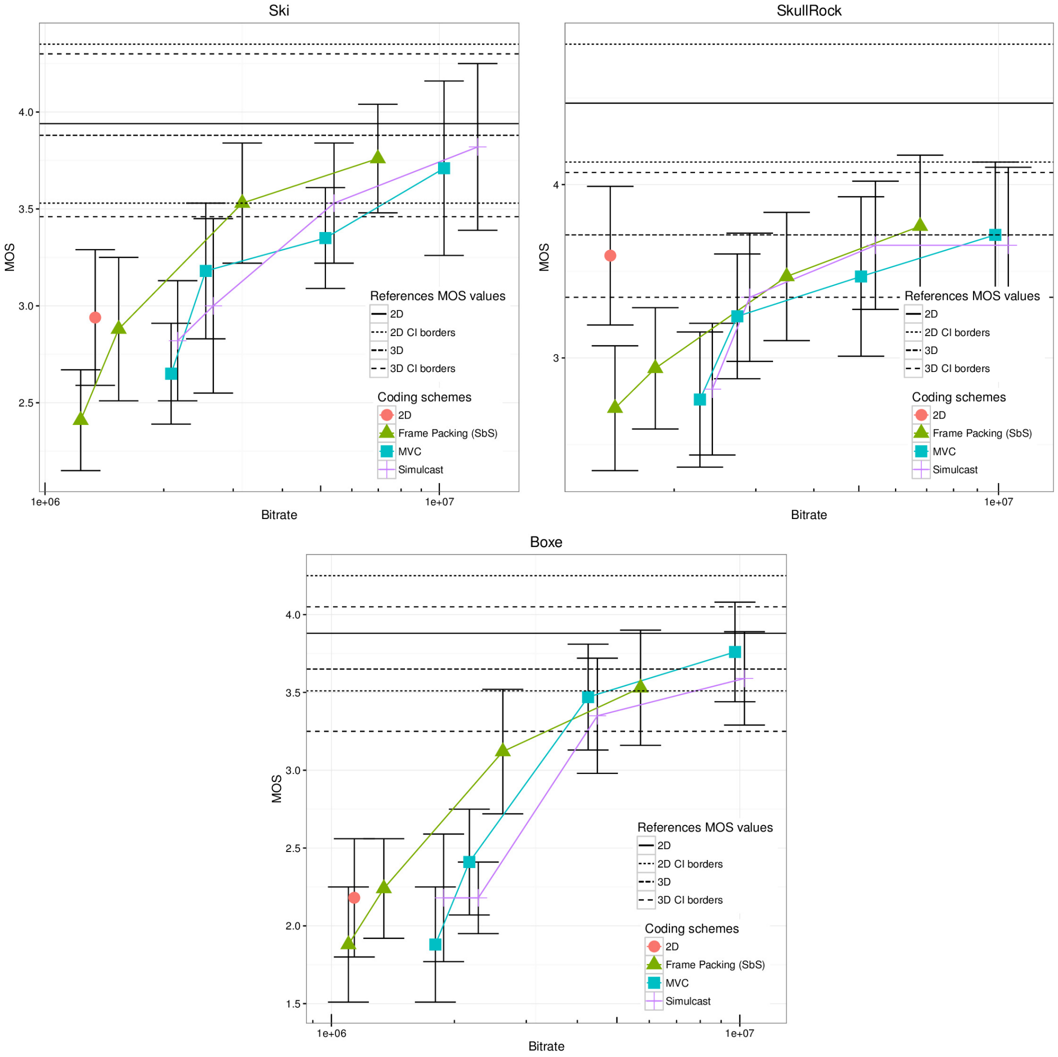


Figure 3.7: Quality per content and per coding scheme (second part of SRCs).

reference. The most likely reason for this result is that it is difficult for the observer to differentiate between high quality contents when the contents are presented sequentially, as it is done in a single stimulus test.

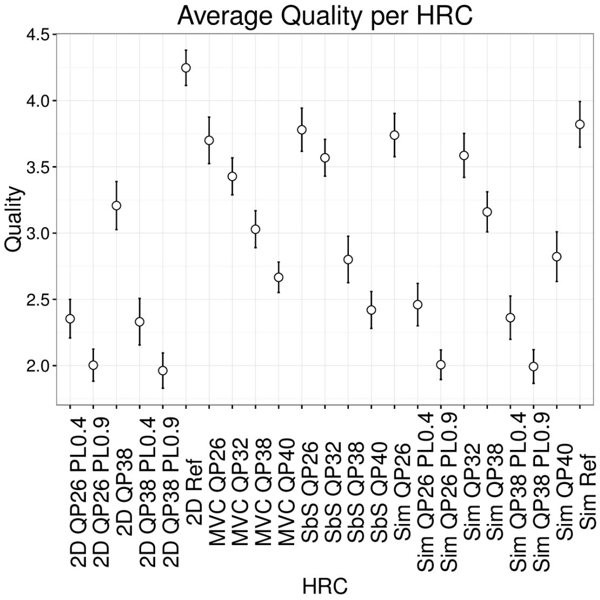


Figure 3.8: QoE rating as a function of the HRC. The notation 2D QP26 PL0.4 indicates a 2D condition encoded with a quantization parameter of 26 and having packet losses introduced with a percentage of 0.4 packet dropped.

| SRC | MOS | Gain compared to Simulcast | Gain compared to MVC |
|---|---|---|---|
| 1 | 3.94 | 31% | 21% |
| 2 | 4.00 | 31% | 28% |
| 3 | 4.12 | 40% | 56% |
| 4 | 4.06 | 57% | 57% |
| 5 | 3.82 | 55% | 48% |
| 6 | 3.65 | 52% | 58% |
| 7 | 3.59 | 55% | 55% |

Table 3.4: Gain in Bitrate of using the SbS representation compared to Simulcast or MVC for a fixed quality level

#### 3.1.1.5 Relation between 3D QoE scores and Visual comfort scores

During the test, test participants were asked to report about visual comfort for each test conditions. Figure 3.9 depicts the relationship between the two scales: QoE and comfort. It can be seen that the test participants used both scales very similarly, and shows a Pearson correlation of 0.88 between them. This very high correlation between the scales can be explained by the fact that test participants must have associated the term of visual comfort with the annoyance they felt with the coding and packet loss impairments, and have then related the two scales. Indeed it was not expected that visual comfort scores would becomes as low, since the test set-up was designed such as the viewing distance was set to enable a comfortable viewing experience. Only transmission impairments were expected to induce visual discomfort. A linear regression between the two factors was performed according to equation 3.1, and show respectively values of 0.77 for $\alpha$ and 0.70 for $\beta$. The value of $\beta$, higher than 0 shows that QoE could be rated lower than comfort, therefore in the term QoE test participants took more factors into account than the visual comfort. This most likely includes the pictorial quality. The slope, $\alpha$, lower than 1, shows that test participants could report video sequences providing a high quality of experience, even though the visual comfort was not optimum. And therefore, they may have taken other factors into account such as the high quality of the picture and perceived depth.

$$Comfort = \alpha \cdot QoE + \beta \tag{3.1}$$

#### 3.1.1.6 Study of 3D QoE in case of a lossy transmission chain

Another important aspect of the experiment was to evaluate the effect of packet loss on the perceived Quality of Experience of 3D videos. The goal were to evaluate how common 2D error concealment strategies perform in case of 3D video, and to compare the quality of 2D and 3D video under packet loss. The evaluation of 3D vs 2D is particularly interesting, since in the 3D case two contradicting factors are involved:

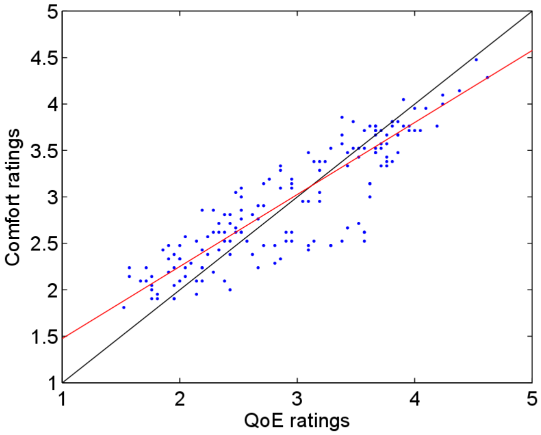


Figure 3.9: Relationship between QoE scores and visual discomfort ratings. The fitting between the two factors, Comfort and QoE rating is represented in red.

1. The binocular suppression theory, which says that if one of the two stereoscopic views has distortions, then the resulting quality can be high, since the quality may mainly depend on the best of the two views, or at least on the average of the quality levels related to each individual view.
2. The binocular rivalries (when one of the two eyes perceives strong artefacts) which induces visual discomfort, affecting the general quality of experience.

Figure 3.10, depicts for every SRC, the Quality of Experience as a function of the error rate. No significant difference could be found between the 2D and 3D conditions, participants only lowly rated the quality of these conditions. An hypothetical reason would be due to the fact that in the experiment, test participants rated in the same test video with only coding distortions, and video with transmission impairments. The fact that transmission impairments provided much stronger distortions than the coding ones may also have compressed the scale and resulted in having test participants rating more severely the video sequences with transmission impairments than the videos having only coding impairments. This have resulted in highly critical ratings for these PVSs with transmission impairments.

#### 3.1.1.7 Conclusion

In this experiment the perceived quality of a current implementation of 3DTV broadcasting, and was evaluated. Its performance was compared with some of the state-of-the-art algorithms. Among the implementations which have been compared, the side-by-side representation seems to be the most efficient way to transmit HD stereoscopic 3D videos, with less bandwidth requirements than Simulcast and MVC using full resolution. These results were in accordance with recent results [101]. The relationship between visual comfort scores and QoE scores was studied. The relation between these two factors was closer than expected, and shows a very high correlation (Pearson correlation of 0.88). Another objective of the experiment was to compare the quality of 2D and 3D videos in case of packet loss. In the test, no significant quality difference between 2D and 3D at a given packet loss rate was observed.
A further result is on the comparison of the scores provided for the 2D and 3D video material: the 3D reference was found to be rated significantly higher than the 2D video material. The results even show cases where 3D is rated lower than the 2D reference. This may be explained by the context of the experiment: ratings video with different coding conditions may have focused the attention of the observer on the image quality aspects.

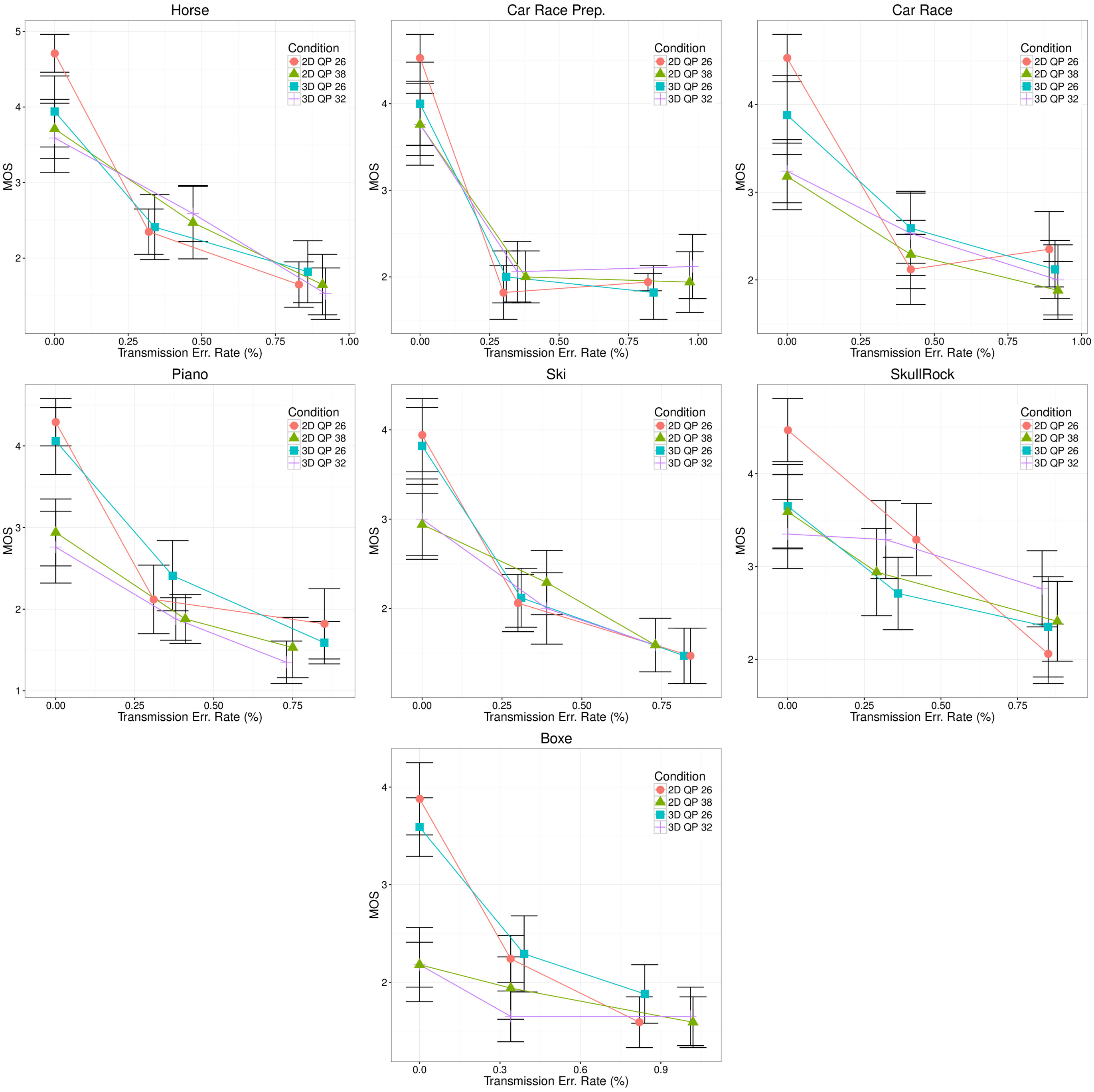


Figure 3.10: Quality per content in function of the percentage of dropped packets

**Listing 3.2: Conclusion on the research questions**

1. & 2. Evaluating differences between 2D and 3D QoE on the reference video, is challenging. By directly asking test participants to rate 3D QoE the differences between 2D and 3D were not observable.
3. For a given bitrate, the Side by Side representation with H.264 encoding appeared to be the most efficient approach. MVC resulted, as expected, in a saving of about 20% of the bitrate for a same subjective score when compared to simulcast.
4. In case of transmission errors, no significant difference between the 3D and 2D ratings were found.
5. The visual discomfort scores were found to be closely correlated to Quality of Experience ratings (Pearson correlation of 0.88). And the relation between the two factors was found, namely $Comfort = 0.77 \cdot QoE + 0.7$.

### 3.1.2 *Further analysis on subjective ratings*

The main issue is is to evaluate, in a subjective manner, the complete 3D experience. Asking directly to rate the quality of experience apparently fails to capture all the different dimensions which are involved in 3D videos. As explained in the state-of-the-art section, different ways have been proposed to evaluate 3D QoE by using alternative evaluation concepts such as *Naturalness*, or *Viewing experience* [2, 6, 7]. Another side of this problematic is the reliability of the subjective methodologies. Existing standards such as ITU-R Recommendation BT.1438 [88] are available and specify methodological settings that target high reliability, but several issues are not addressed as listed by Chen [92] to ensure stable results across laboratories. In the meantime standards have been updated to tackle the question of how 3D QoE should be evaluated and in which environment. This can be found in ITU-R Recommendation BT.2021 [102]. To investigate these issues in this thesis two main analysis was performed based on the previously described subjective experiment.

**Listing 3.3: Research questions**

1. How do subjective scores compare from one laboratory to another? Hence, how stable is a subjective test ?
2. How do test participants understand and rate on the different scales they are asked to use ?

#### 3.1.2.1 Inter-laboratory comparison

The first analysis considered here is the comparison of test results with results obtained by different laboratories. Two other laboratories (L1: ACREO, L2: IRCCyN) have conducted experiments with similar conditions as our experiments described in section 3.1.1. The comparison of their experimental results was part of the analysis presented in [100]. In our experiment we used different SRCs. Only 9 HRC were common between the tests (1,2,5,6,10,11,12,18,21). Another difference between our tests was that people saw video with transmission impairment in our test so that the ranges of degradation types and quality were clearly different. In addition, the methodology used for evaluating the visual discomfort in the test is different: in our test we used ACR with a 5 grade scale, the other tests also used ACR-HR but the vocabulary used indicated a comparison with the 2D viewing (e.g. the 3D presentation was: much more comfortable than 2D, more comfortable than 2D, as comfortable as 2D...).
Figure 3.11 depicts a direct comparison between the labs quality test results on identical HRCs. As the SRCs were different, a direct comparison of the score is not possible and only an overall condition MOS can be compared. Due to this average, it is not possible to compare the absolute MOS values corresponding to each HRC, however it can be checked whether a linear relationship and not a more complex non-linear relationship exists between the results of our experiment and results from the literature. With this goal in mind, a linear model model applied to compare the results

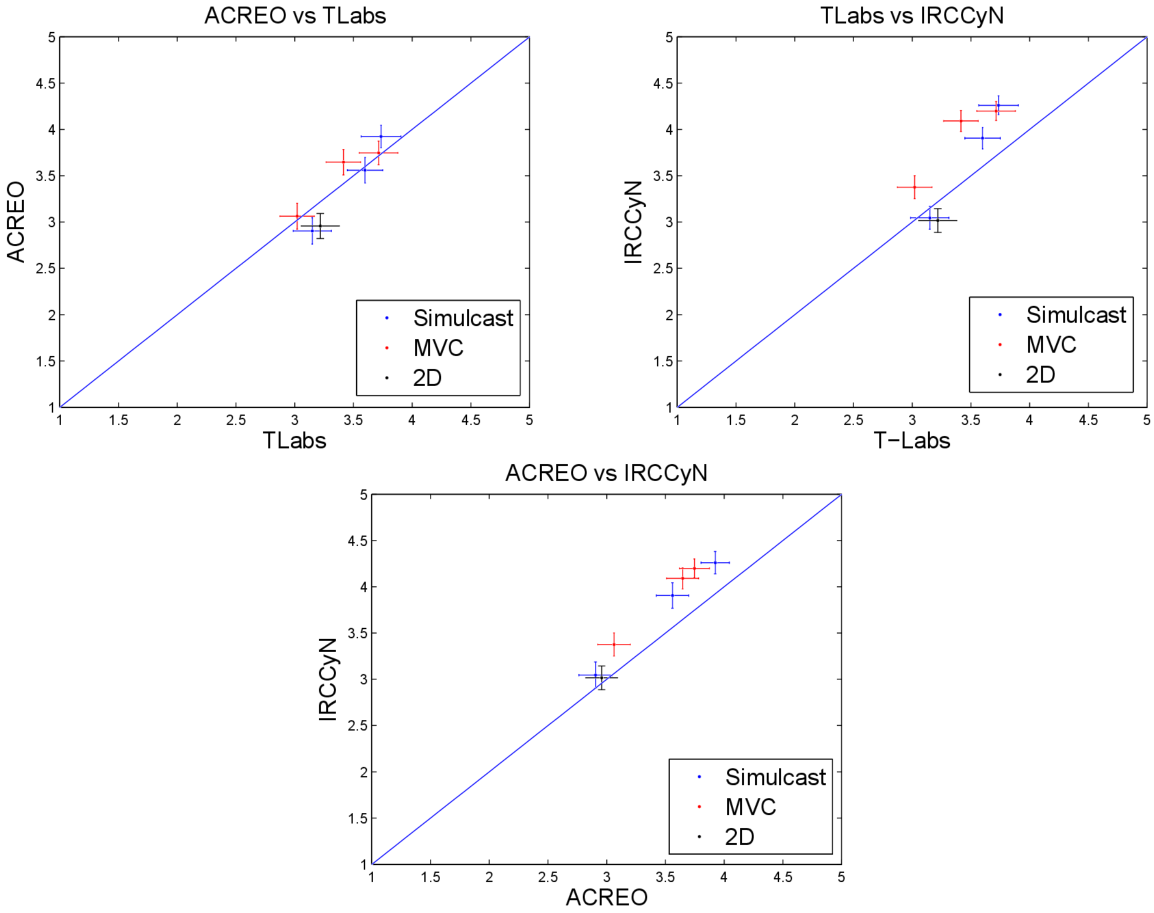


Figure 3.11: Comparison of quality evaluation between laboratories

between the labs. The Pearson correlation is 0.84 between our test and L1, 0.71 with L2 and 0.97 between L1 and L2. The correlation is high considering the different assumptions and averages made across the different SRC.

The numbers of similar test conditions between the different laboratories is too small to draw strong conclusions. However these results go into the same direction as the study performed by Wang at al. [100] and more recently Barkowsky et al. [103] on inter-laboratory result stability when the evaluation of video sequences with different coding conditions is considered.

#### 3.1.2.2 Study of scale usage

Another interesting point was found when addressing the visual discomfort evaluation. Figure 3.12 depicts a direct comparison of the quality and discomfort ratings. From this figure it becomes apparent that observers have answered differently in case of the visual discomfort scale: there are subjects who have rated quality and discomfort in a similar fashion and others who did not. To further analyze these variations, the observers were classified into different classes. It can be stated that this problem was not specific to our test and was also visible in the tests results of L1 and L2 described in the state-of-the-art section. It appears that visual discomfort is a difficult concept and not all observers understand the scale in an identical way.

The different patterns of answers observed leads to the following classification:

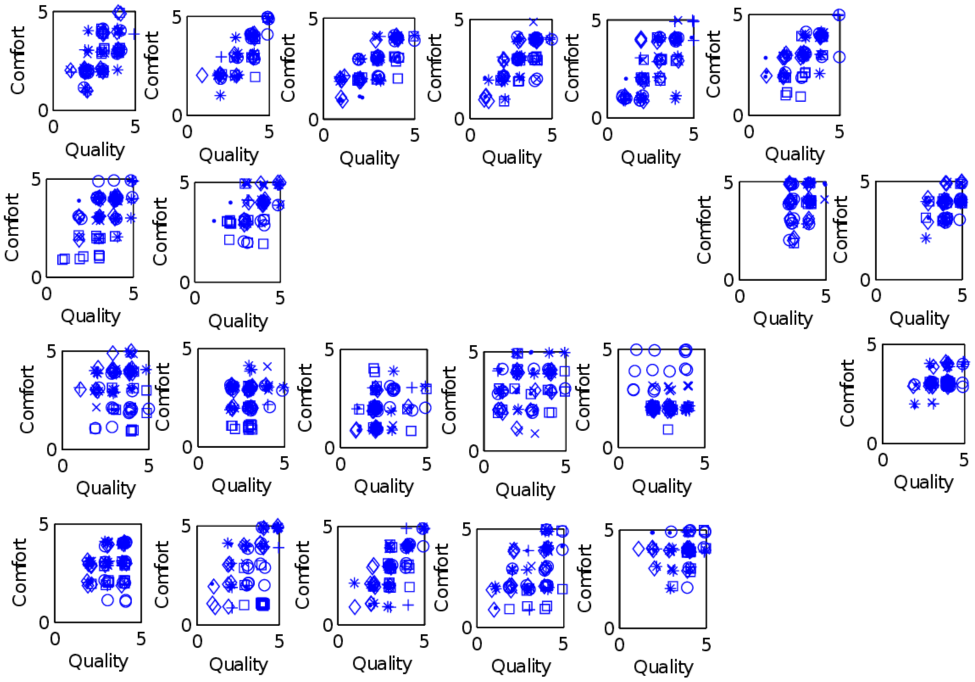


Figure 3.12: Different types of answers between observers: each scatter plot represents a type of observer in terms of her answers. The scatter plot are grouped showing different patterns of scores.

1. Observers who answered with a clear linear relation between discomfort and the quality scale. These participants either have considered a direct relation between discomfort and quality, or were simply not able to distinguish between the two concepts.
2. Observers who completely covered the space with different quality–discomfort rating value pairs for different HRCs. This group apparently considered quality and discomfort to not necessarily be related, for example, when the discomfort is mainly due to the content.
3. Observers showing an answer pattern in the shape of a triangular matrix: the value of discomfort is between [1, CMax] with CMax being a function linearly dependent on the quality. These observers apparently have considered a relation of implication between comfort and quality: a high discomfort leads to a low-quality video, but the reverse was not necessarily true: low quality video could be due to degradations that are not related to discomfort.

Figure 3.13 depicts the distribution of the ratings correlation across the different participants. Using k-means, the participants were clustered into two groups based on the Pearson correlation between the two rating scales: quality and comfort. It should be mentioned that previously three classes were mentioned. However the Pearson correlation may not be a good indicator to characterize a triangular matrix shape of replies. This is why, for the sake of simplicity only two classes are considered in the following analysis. Based on the kmeans two clusters having the means of 0.327 and 0.742, and the respective within clusters sum of squares of 0.118 and 0.136 were found. The within sum of squares being relatively small indicates that two clusters could be formed according to the Pearson correlation indicator.
Based on these clusters, a non-parametric Wilcoxon rank sum test was applied and show that these two groups significantly rated differently on the comfort scale (W=1338600, p=0.025), but did not on the quality scale (W=1426700, p=0.323).

Figure 3.15 depicts the average ratings of visual discomfort as a function of the HRCs. Since the observers rated discomfort differently, the following analysis is performed by groups. To create these classes, the correlation of each observer between their quality and visual discomfort ratings were determined. Then, a k-means analysis of these corre-

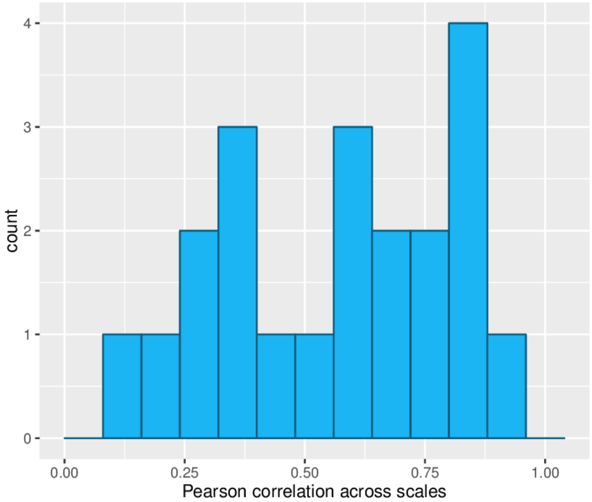


Figure 3.13: Distribution of the Pearson correlation values between the Quality and Comfort scales for the different participants.

lation values was performed and used to divide the observers into the ones who have clearly related visual discomfort and quality (class with R between 0.63 and 0.92 with an average of 0.74) and the ones who did not (R between 0.13 and 0.51 with an average of 0.33).

From the two curves we can see that observers who have not necessarily linked quality and discomfort gave more constant ratings of visual discomfort than the other class of observers. However, also some of the 2D sequences were rated as uncomfortable by these users, which is an unexpected behavior.
When comparing the results with the L1 and L2 tests, it can be stated that there, too, a high variation of discomfort ratings could be observed, although discomfort was rated relative to the 2D version. Hence, the tests underline the difficulty of judging discomfort of 3D video.

Engelke [104] performed a similar study and concluded that test participants do not necessarily understand all the different scales they have to use in the same manner. It then may be difficult for the test participant to rate complex notions such as *Naturalness* or *Visual experience* as proposed by Seuntiëns [2] since these tests show that test participants already have difficulties to understand evaluation concepts such as *Quality of Experience* and *Visual discomfort*. Therefore, the next section of this thesis will focus on subjective evaluation methods which are simple for test participants and enable to quantify the added value of 3D.

### 3.1.3 Conclusion

In this subsection, the question of subjective test reliability was addressed. The results of the test shows that tests from one laboratory to another appear consistent when test participants are asked to rate QoE of videos encoded with different conditions. However, the results show that test participants may have difficulties to clearly understand the scales they have to use, and variation from test participants to another can be seen. This shows, as also reported by Engelke [104], that too high-level evaluation concepts may be difficult for the test participants, and therefore there is a need to design easier experiments for 3D video sequences evaluation when lots of dimensions are involved.

**Listing 3.4: Conclusion on the research questions**

```
1. Test participants use the scales differently, therefore there is a need to simplify
   the task of the test participants when lots of dimensions are involved in the
   evaluation.
```

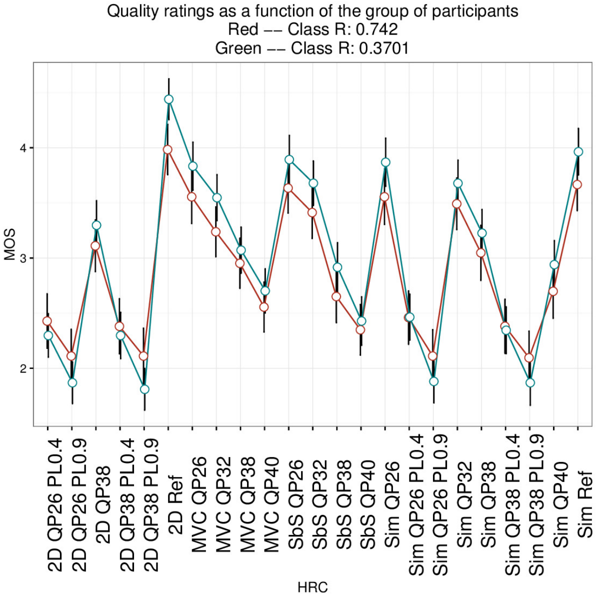


Figure 3.14: Quality ratings in function of the HRCsfor the two classes of observers: the ones with a low correlation between quality and discomfort and the one with high correlation between quality and discomfort

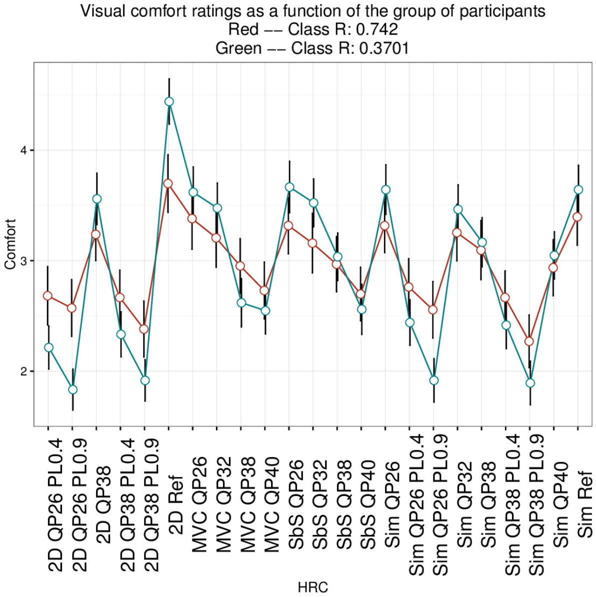


Figure 3.15: Visual discomfort in function of the HRCs for the two classes of observers: the ones with a low correlation between quality and discomfort and the one with high correlation between quality and discomfort

### *3.1.4 Performance of instrumental measurement for 3D QoE prediction*

Considering the similarity between the QoE scores obtained from test participants when asked to rate 3D and 2D QoE, it appears that test participants do not fully take all the factors covered by the concept of 3D QoE into account. Based on this result, different questions were raised:

**Listing 3.5: Research questions**

```
1. If the rating obtained for 3D does not strongly differ from 2D, how do 2D quality
    prediction algorithms perform when predicting 3D quality?
2. How do two standardized algorithms perform when applied to 3D contents ?
```

#### 3.1.4.1 Selection of the conditions

The idea is to emulate the real signal chain in a 3DTV broadcasting solution and study how 2D video quality prediction algorithms perform for the evaluation of 3D QoE. Therefore the test design consisted of a live hardware encoder which was fed by a hardware playback server for uncompressed playback. The encoder's output was sent to an IPTV server and finally the signals were streamed to a test set-top box. The HDMI output of that set-top box was captured and recorded on a MacPro equipped with a video acquisition interface card. The sequences were then stored using the Apple ProRes 422 (hq) codec at a bit rate of around 180 mbit/s. The setup of the recording is depicted in Figure 3.16.

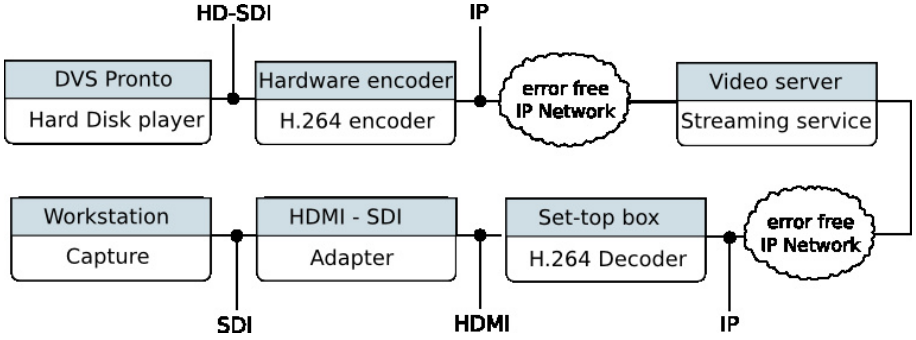


Figure 3.16: Processing chain for the creation of PVSs

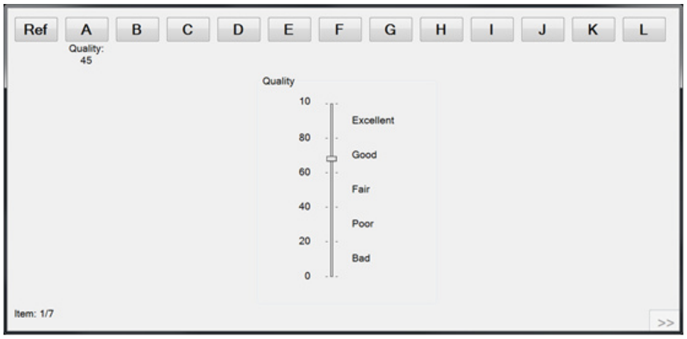
In the next step, the sequences were edited by means of Final Cut Pro without changing the format of the recorded clips to extract the video sequences selected for evaluation after stabilization of the encoder. The experimental condition consisted of using the hardware encoder at ten different bit rate values (5, 7.5, 10, 12, 14, 16, 18, 20, 22, 24Mbps) and a software encoder at one bit rate value (7.5Mbps) used for comparison. Seven different source signals were chosen, the sequences had different spatial, temporal and depth complexity. A short description of the sequences is provided in Table 3.5

#### 3.1.4.2 Subjective evaluation method

The subjective test methodology SAMVIQ was chosen [105]. This methodology consists in presenting several sets of video sequences to the observers. In each set, several sequences are presented. These sequences contain the same source signal but with different processing. The observers can choose a video from the proposed sequences within the set, watch it and rate it. One of the sequences is clearly identified as the reference, and one is a hidden reference. The observers can repeatedly watch each sequence and adjust the respective rating. After having watched and rated

| Content Name | Description |
|---|---|
| Bear | Sequence from animation movie. Complex motion: lots of particles, and strong movement; Lots of high frequency texture. 3D with pop-out effect. |
| Fans | Soccer fans with many small details. Complex motion: fans are moving, shaking flags. |
| Horse | Sequence with strong texture and limited motion: horse standing and starting running. |
| Interview | Sequence with two persons interviewed. The background is composed of trees moving in the wind. Limited motion. Some pop-out effect is visible: the arm of the persons comes out the screen. |
| Match | Football match, lots of high frequency texture on the grass. Fast motion. |
| Piano | Sequence with low spatial and temporal complexity. Piano player sitting in front of the piano and standing up. |
| Sea | Sequence with sea water during a storm. Lots of high frequency textures. Complex but slow motion. |

Table 3.5: 3D video content characteristics



Quality:
45

Item: 1/7

Figure 3.17: Subjective experiment interface used for the evaluation of the video sequences

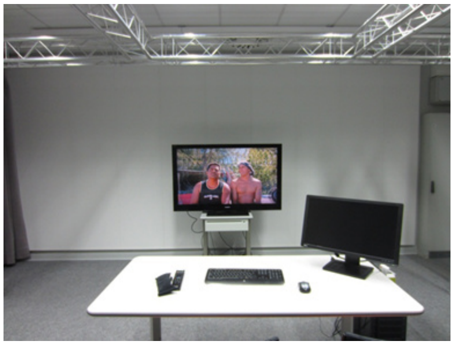


Figure 3.18: Setup of the laboratory environment

all videos of one set he can continue to the next one. The choice of SAMVIQ was motivated by the fact that this methodology gives the ability to compare different video signals to an explicit reference which helps the observers to evaluate the quality of a specific processed sequence. The eventual repetitions provide the ability to adjust the rating which is particularly useful the present the case of this study since many conditions had similarly high quality. Providing an explicit reference and a way to readjust a given score can help the subject to evaluate the different sequences. This is confirmed in previous studies which show that SAMVIQ can be more stable than ACR if the observer uses the replay feature [106].

The test condition was set in accordance to ITU-R Recommendation BT.500-12 [98]. The viewing distance was 3 times the height of the screen (3H). The playback computer was a Pentium Core i7 PC with a graphic card which had an HDMI output. The Stereoscopic Player [107] which was used for playback of all videos was running in full-screen mode on the secondary display. The 3D sequences were displayed on a commercial Sony 52” TV screen using shutter glasses, the interface for the subjective testing was presented on another PC display connected to the same computer (see Figure 3.18). The test subjects were persons involved in research and development, but no professionals working on topics such as TV editing or production on a daily basis. 19 subjects were participating. The task was demanding: finding small differences in steps of 2 mbits/s between 10 and 24 Mbit/s.

### 3.1.4.3 Subjective evaluation results

The subjective scores for each source sequence are depicted in Figure 3.19. As a first outcome it is visible that with the same set of parameters and at the same bitrate the hardware encoder performs better than the software encoder. The differences are statistically significant at a 95% confidence level using the student-t test for three out of the seven contents (Fans, Match, Sea).

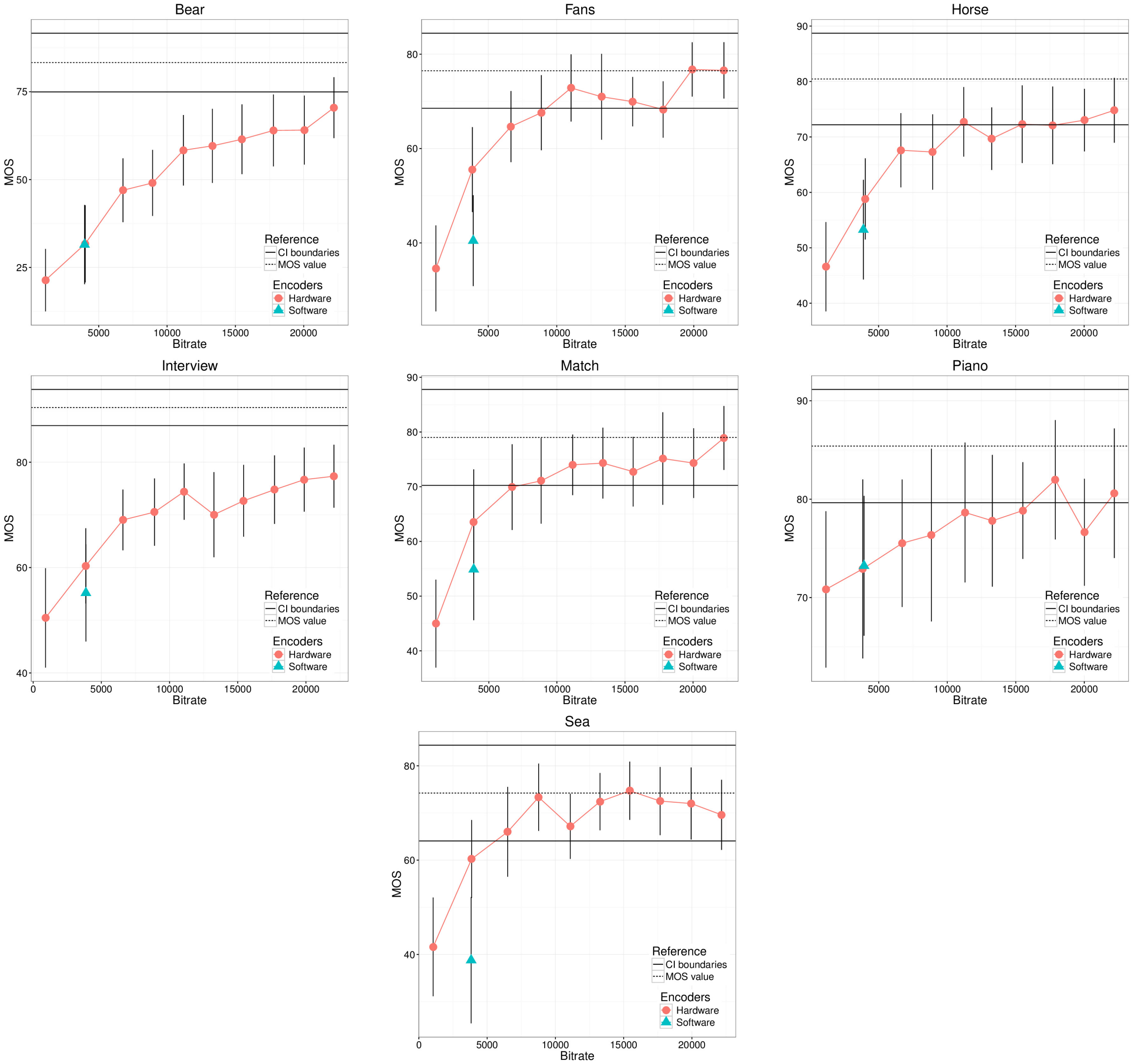
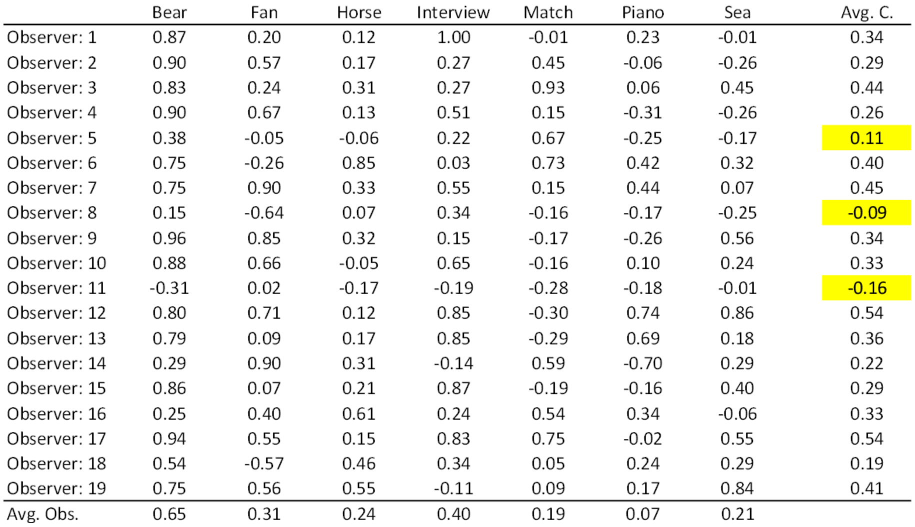


Figure 3.19: Subjective quality score per content as a function of the bitrate in kbps

As depicted in Figure 3.19, the confidence intervals are quite large. This is most likely due to the difficulty of the task asked from the observers: many conditions had high quality hence it was difficult for the observers to be able to give accurate absolute quality ratings. However since the SAMVIQ methodology was employed, observers had the opportunity to compare each sequence to others ones. Comparing the sequences gave them the ability to reveal their preference for one preferred sequence compared to another one in terms of compression artifacts. Even though it was hard for them to give absolute subjective scores, in most cases test subjects were able to provide relative ratings. The Spearman Rank Order Correlation of each individual observer with other is depicted in Table 3.6. To build this matrix the coding conditions with the hardware encoder were considered, and it is believed that increasing the bitrate will decrease the value of the quantification parameters and therefore increase the quality. Subjective scores should then follow this evolution. If there would have always been a clear improvement of the quality with increasing bitrate, the observers might have obtained a Spearman Rank Order Correlation of 1. But since the task was demanding, the observers did not provide that accuracy. Based on this analysis three different observers appear to be outliers (They are



| | Bear | Fan | Horse | Interview | Match | Piano | Sea | Avg. C. |
|---|---|---|---|---|---|---|---|---|
| Observer: 1 | 0.87 | 0.20 | 0.12 | 1.00 | -0.01 | 0.23 | -0.01 | 0.34 |
| Observer: 2 | 0.90 | 0.57 | 0.17 | 0.27 | 0.45 | -0.06 | -0.26 | 0.29 |
| Observer: 3 | 0.83 | 0.24 | 0.31 | 0.27 | 0.93 | 0.06 | 0.45 | 0.44 |
| Observer: 4 | 0.90 | 0.67 | 0.13 | 0.51 | 0.15 | -0.31 | -0.26 | 0.26 |
| Observer: 5 | 0.38 | -0.05 | -0.06 | 0.22 | 0.67 | -0.25 | -0.17 | 0.11 |
| Observer: 6 | 0.75 | -0.26 | 0.85 | 0.03 | 0.73 | 0.42 | 0.32 | 0.40 |
| Observer: 7 | 0.75 | 0.90 | 0.33 | 0.55 | 0.15 | 0.44 | 0.07 | 0.45 |
| Observer: 8 | 0.15 | -0.64 | 0.07 | 0.34 | -0.16 | -0.17 | -0.25 | -0.09 |
| Observer: 9 | 0.96 | 0.85 | 0.32 | 0.15 | -0.17 | -0.26 | 0.56 | 0.34 |
| Observer: 10 | 0.88 | 0.66 | -0.05 | 0.65 | -0.16 | 0.10 | 0.24 | 0.33 |
| Observer: 11 | -0.31 | 0.02 | -0.17 | -0.19 | -0.28 | -0.18 | -0.01 | -0.16 |
| Observer: 12 | 0.80 | 0.71 | 0.12 | 0.85 | -0.30 | 0.74 | 0.86 | 0.54 |
| Observer: 13 | 0.79 | 0.09 | 0.17 | 0.85 | -0.29 | 0.69 | 0.18 | 0.36 |
| Observer: 14 | 0.29 | 0.90 | 0.31 | -0.14 | 0.59 | -0.70 | 0.29 | 0.22 |
| Observer: 15 | 0.86 | 0.07 | 0.21 | 0.87 | -0.19 | -0.16 | 0.40 | 0.29 |
| Observer: 16 | 0.25 | 0.40 | 0.61 | 0.24 | 0.54 | 0.34 | -0.06 | 0.33 |
| Observer: 17 | 0.94 | 0.55 | 0.15 | 0.83 | 0.75 | -0.02 | 0.55 | 0.54 |
| Observer: 18 | 0.54 | -0.57 | 0.46 | 0.34 | 0.05 | 0.24 | 0.29 | 0.19 |
| Observer: 19 | 0.75 | 0.56 | 0.55 | -0.11 | 0.09 | 0.17 | 0.84 | 0.41 |
| Avg. Obs. | 0.65 | 0.31 | 0.24 | 0.40 | 0.19 | 0.07 | 0.21 | |

Table 3.6: Spearman rank correlation of each individual observer

observers 5, 8, 11, 18 visible in Table 3.6). It may be arguable to average the Spearman correlation across the video sequences as there are large variations of correlation values for the different source sequences. This is why a second analysis was performed to cross check the inter-participant agreement, and participant removal. The scores provided by each participants was directly compared to the scores of the other participants in term of Pearson correlation. Figure 3.20 depicts the inter-participant agreement using this other criteria. As in the previous analysis, participants 11 and 8 appears as clear outliers. Participant 5 do not also provide a high agreement with the other participants. However, participant 18 shows a stronger agreement with other participants such as 4, 7, 9, 14. This is why it was decided not to reject this participant. In this second analysis, the participant 19 also shows a lower agreement than the other participants. However, in the ranking analysis, this participant perform along the most consistent ones. Therefore, it was not rejected as well. This participants screening have resulted in the rejection of only three participants: 5, 8 and 11.

A result of this test is a method to determine the bitrate value from which an increase of bitrate will not provide an increase of quality perceivable by the observers. Considering the size of the confidence intervals, it is proposed to use the fact that using SAMVIQ, even though observers had difficulties to agree on an absolute quality value for a sequence they were at least able to order the sequences. Then, it is possible to check the monotony of the quality scores; which should be in accordance with the increase of bitrate. The point from which this agreement is broken, should be then assumed to be the point where observers were not able anymore to see the difference between the quality of the sequences. The bitrate threshold is then obtained at this specific value. Table 3.7 provides, for each observer and for each content, the bitrate threshold determined as proposed previously. It is then proposed, for each content, to take the average of the bitrate value obtained for each observer as the expected threshold.

#### 3.1.4.4 Prediction of 3D QoE

To evaluate the quality of broadcasted IPTV, another typical approach could be the use of instrumental models. It is proposed to evaluate the accuracy of two standardized models in evaluating the quality of 3D video sequences: VQM and VQuad. The models were run on video sequences with the side-by-side representation. Figure 3.22 depicts the performance of VQM on the previously presented database. The model achieves good performance with a Pearson correlation of 0.8947 and an RMSE of 5.4 (after a linear mapping to a 0-100 scale: MOSe = -119.6 * VQM + 86.92). It should be noted that the subjective scores of the video sequences mainly lie between 50 and 80, which may result in a high value of Pearson correlation. Figure 3.21 depicts the performance of VQuad on the proposed database.

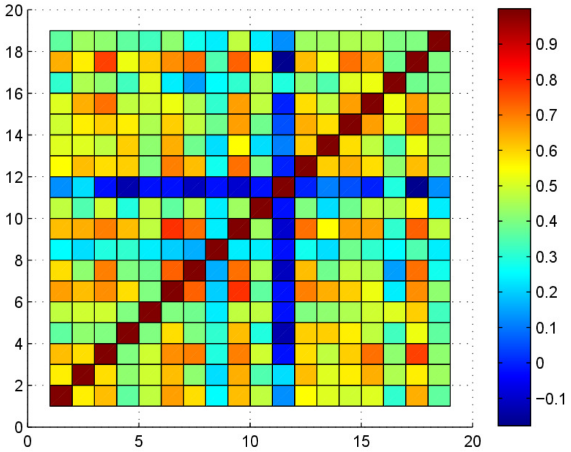
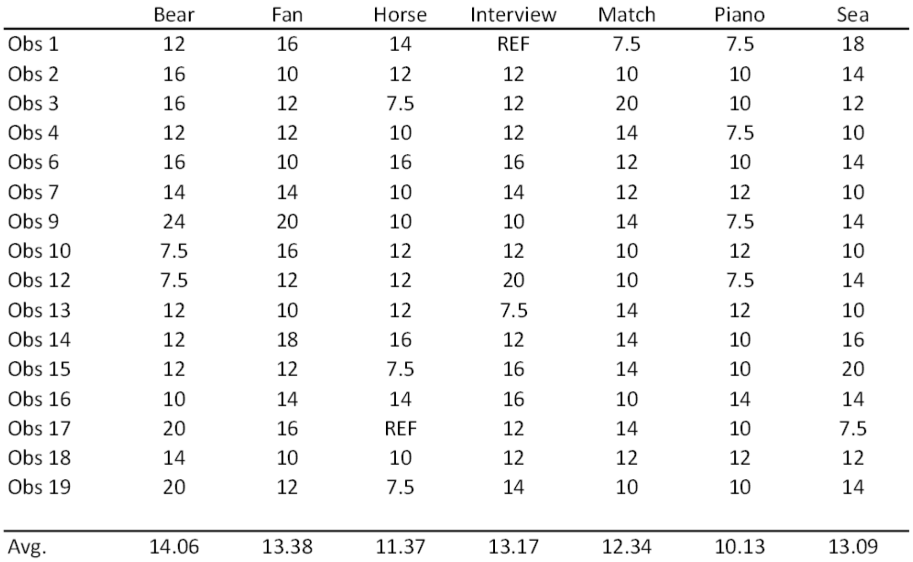


Figure 3.20: Matrix of cross Pearson-correlation between participants ratings



| | Bear | Fan | Horse | Interview | Match | Piano | Sea |
|---|---|---|---|---|---|---|---|
| Obs 1 | 12 | 16 | 14 | REF | 7.5 | 7.5 | 18 |
| Obs 2 | 16 | 10 | 12 | 12 | 10 | 10 | 14 |
| Obs 3 | 16 | 12 | 7.5 | 12 | 20 | 10 | 12 |
| Obs 4 | 12 | 12 | 10 | 12 | 14 | 7.5 | 10 |
| Obs 6 | 16 | 10 | 16 | 16 | 12 | 10 | 14 |
| Obs 7 | 14 | 14 | 10 | 14 | 12 | 12 | 10 |
| Obs 9 | 24 | 20 | 10 | 10 | 14 | 7.5 | 14 |
| Obs 10 | 7.5 | 16 | 12 | 12 | 10 | 12 | 10 |
| Obs 12 | 7.5 | 12 | 12 | 20 | 10 | 7.5 | 14 |
| Obs 13 | 12 | 10 | 12 | 7.5 | 14 | 12 | 10 |
| Obs 14 | 12 | 18 | 16 | 12 | 14 | 10 | 16 |
| Obs 15 | 12 | 12 | 7.5 | 16 | 14 | 10 | 20 |
| Obs 16 | 10 | 14 | 14 | 16 | 10 | 14 | 14 |
| Obs 17 | 20 | 16 | REF | 12 | 14 | 10 | 7.5 |
| Obs 18 | 14 | 10 | 10 | 12 | 12 | 12 | 12 |
| Obs 19 | 20 | 12 | 7.5 | 14 | 10 | 10 | 14 |
| Avg. | 14.06 | 13.38 | 11.37 | 13.17 | 12.34 | 10.13 | 13.09 |

Table 3.7: Bitrate threshold for perceived quality difference in mbps

This second model achieves lower performance on the studied database: it shows a Pearson correlation of 0.7586 and an RMSE of 8.2 (after a linear mapping to a 100 scale: MOSe = 17.49 * VQuad + 4.628). It should be taken into account that the VQuad model is able to handle video sequences with packet losses, which VQM is not. Therefore, we can argue this may have an influence on the performance. When only high-quality sequences are considered, VQM is more appropriate to evaluate the quality of encoded sequences before transmission, VQuad would be more suited for the evaluation of video sequences at the end of the transmission chain. Since transmission impairment is a dominant artefact compared to coding, the development of VQuad may have been less focused on transmission-error-free sequence. Then, the accuracy is lower for the specific scope of our study. VQM here seems to be more suited for such transmission-error-free test.

Considering the performance of the VQM model, a second aspect of this study is to attempt to determine the bitrate corresponding to the quality saturation using an objective method such as presented for the subjective test in

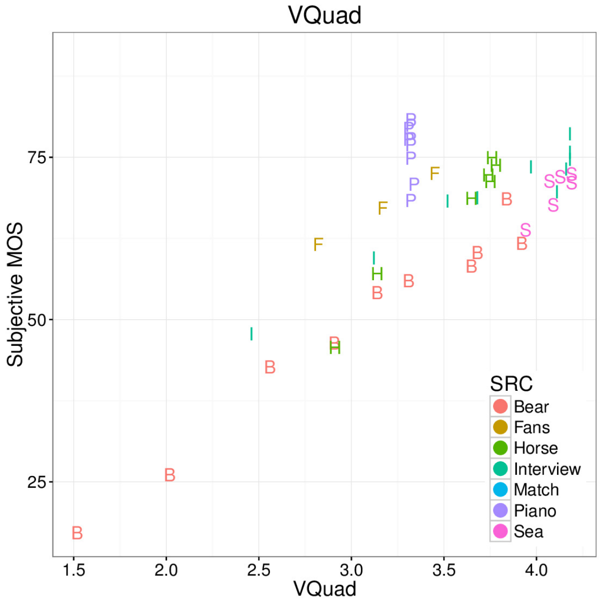


Figure 3.21: Results of the VQuad model (The first letter is used as a marker to identify the SRC)

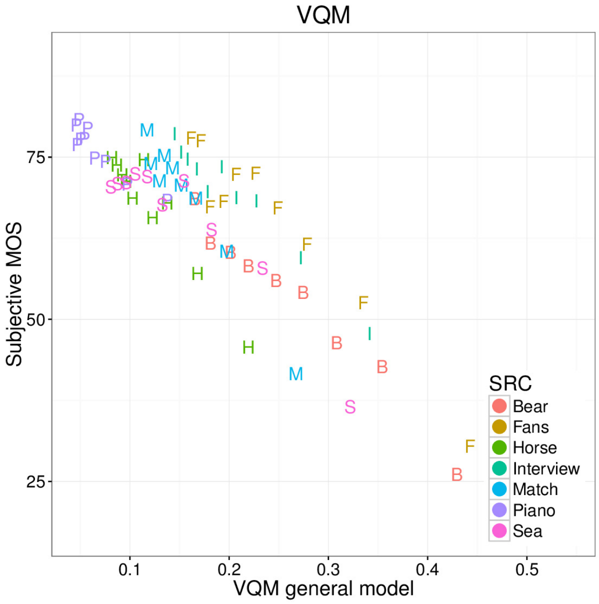


Figure 3.22: Results of the VQM general model (The first letter is used as a marker to identify the SRC)

the previous section. Figure 3.23 depicts, for each content, the subjective and objective quality evaluation as a function of the logarithm of the bitrate. It can be noticed that some sequences may still increase their quality outside of the evaluation interval (strong effect for Bear, Fan; less for Interview and Sea, and only a small effect for Horse, Match and Piano).

In the following evaluation, a different method is proposed to identify the quality saturation. The VQM algorithm was only used as an example of an objective metric. The idea is based on the observation that at very high bitrates the quality of the video tends to converge and once a certain quality level is reached an increase of bitrate does not provide significant increases of visual quality. In the specific instantiation of this study, according to the fitting, the maximum visual quality is reached at 89.5 MOS (but could be different in another experiment). It may be anticipated that the subjects are not capable of appreciating the quality gain related to a video that is above a certain bitrate threshold, for example 95% of this maximum quality. In that case, a certain bitrate can be saved by identifying, using the VQM algorithm, which bitrate corresponds to 95% of the maximum quality. In this evaluation, the value of 24Mbps has been used as the maximum quality prediction. A linear fitting has been performed on the log-bitrate/quality scale and the 95% as well as the 90% quality points has been extracted. The results are presented in Table **??**. The equivalency of these results in terms of subjective score is given in Table 3.9

These results provide a range of bitrates which matches to the subjective bitrate threshold determined in the previous section. This may provide an instrumental method to estimate a range of bitrates around the saturation point.

It should be noted that for the piano sequence, most observers inverted their preference already at very low bitrate, mostly at the second or third bitrate step. In this case, the objective method provides a value which is even lower than the smallest possible value obtained from the subjective experiment (7.5Mbps). Considering the subjective experiment method, the objective metric might even provide an estimation in this particular case that is close to a more generally valid threshold level.

### 3.1.4.5 Sum up

There are two main outcomes of this study: first, a method was presented for determining the saturation point of 3D QoE when increasing the bitrate. Due to the difficulty of performing a subjective experiment requiring the comparison

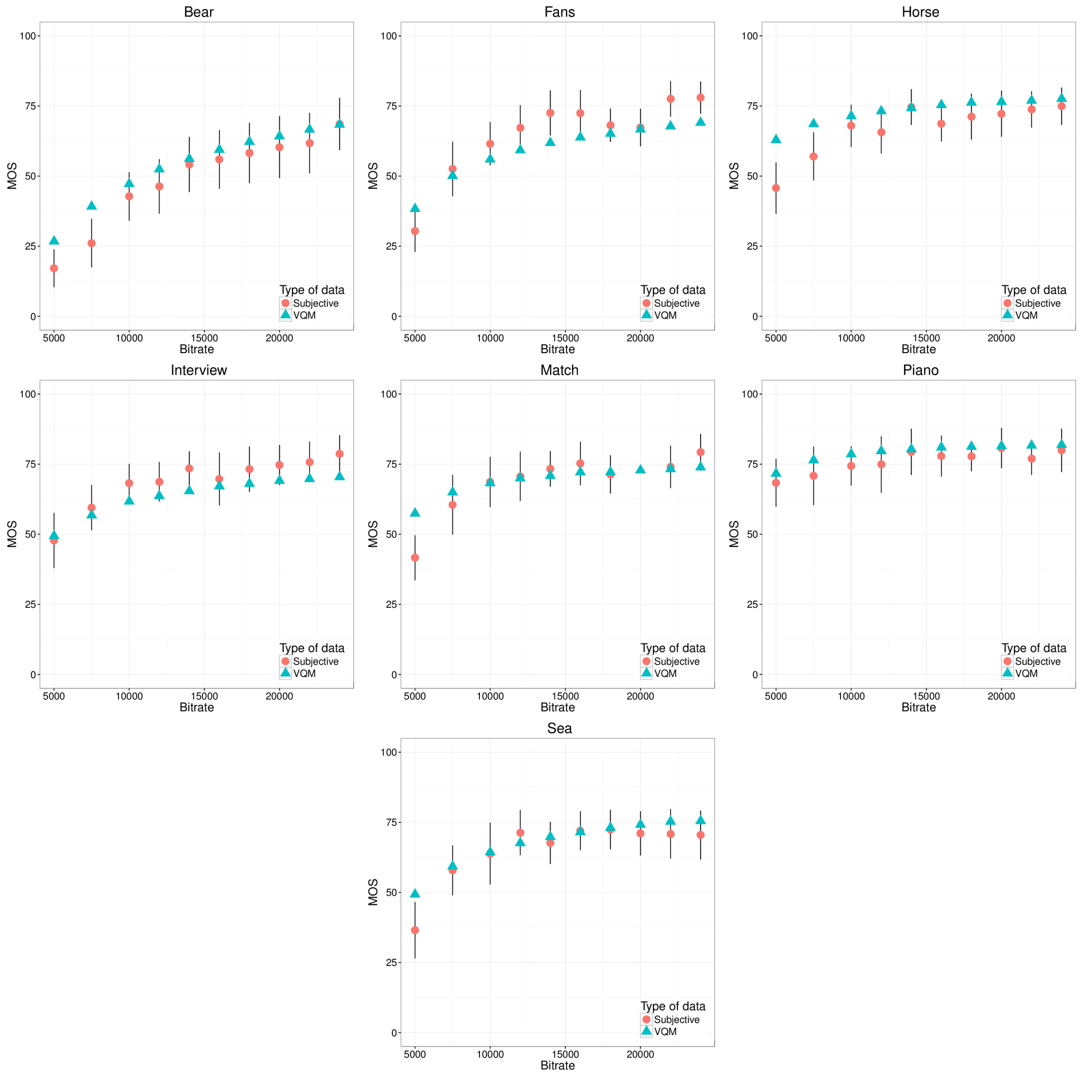


Figure 3.23: Objective and subjective video quality as a function of the logarithm of the bitrate

of many similar high-quality sequences, it was proposed to use the ranking obtained by the SAMVIQ methodology to determine this threshold. The second and main result was the comparison of two standardized objective models (VQM and VQuad) for estimating the quality of the 3D video sequences. The VQM model has shown good performance

| Content Name | 95% Max Quality | 90% Max Quality |
|---|---|---|
| Bear | 17.3Mbps | 16.3Mbps |
| Fans | 18Mbps | 14.4Mbps |
| Horse | 14.6Mbps | 8.6Mbps |
| Interview | 16.5Mbps | 11.7Mbps |
| Match | 12.3Mbps | 8.8Mbps |
| Piano | 5.9Mbps | 5.2Mbps |
| Sea | 16.7Mbps | 12.3Mbps |

Table 3.8: Bitrate value from which 90% and 95% of the maximum objective quality is achieved

| Content Name | 95% Max Quality | 90% Max Quality | Subj. threshold |
|---|---|---|---|
| Bear | 66.96 | 63.43 | 58.20 |
| Fans | 72.49 | 68.67 | 73.74 |
| Horse | 70.92 | 67.19 | 67.45 |
| Interview | 73.49 | 69.61 | 72.69 |
| Match | 73.48 | 69.61 | 71.55 |
| Piano | 76.48 | 72.45 | 75.45 |
| Sea | 65.95 | 62.47 | 69.71 |

Table 3.9: Subjective values corresponding to 90% and 95% of the maximum objective quality is achieved and subjective score corresponding to the bitrate threshold defined subjectively

on the proposed database and seems to be appropriate for tuning the settings of an encoder. As a last result, a way to determine an interval of bitrate around the quality saturation point using an objective measurement method was described. These results are consistent with the work of Benoit [108].

The ability of instrumental algorithms for predicting 3D QoE confirms that evaluating the overall 3D QoE can be challenging: these algorithms without taking into account any information about the depth or visual comfort, but only texture quality are able to predict the scores obtained via subjective ratings. This extends the results from Benoit [108] obtained on still images, and similarly concludes that a good 2D prediction algorithm can be applied to evaluate 3D quality in case of video encoded with traditional coding algorithms such as H.264.

A last contribution from this experiment was to provide a method to estimate the quality saturation using prediction algorithm such as VQM.

**Listing 3.6: Conclusion on the research questions**

1. By applying a linear fitting on 2D quality prediction algorithms, it is possible to predict 3D video quality for coding degradation using 2D video coding algorithms.
2. The comparison of VQM and VQuad, shows that in the particular context of this study VQM performed better than VQM.

## 3.2 Revealing the added value of 3D over 2D

In order to reveal the differences of user experience of test participants while watching 3D video sequences compared to 2D video sequences, alternative evaluation schemes have been considered by Seuntiëns. They address other dimensions like naturalness or immersion, or investigate specific factors like depth perception or eye-strain [2], which provides some insight into isolated factors of the general QoE, but not the overall QoE. In the present study, as global measure of QoE, the subjective preference has been considered. It is believed that when subjects are asked for preference between two videos, they may consider all factors (picture quality, both depth quantity and depth quality, visual discomfort and probably other factors) to take the decision which of two versions of a sequence they prefer. This way, the entire multidimensionality of 3D QoE is considered. Missing factors of 2D video QoE when evaluating it using ACR were shown by Belmudez [109], where another multidimensional question was studied. Here, image size and image resolution were compared in terms of quality ratings, one using ACR, one using paired comparison (PC). Results showed that the two test methods do not provide the same results: using ACR, observers give higher QoE ratings for images at their native resolution; using PC, observers prefer larger images obtained after upscaling. The results are different, and show that using the ACR methodology observers only judge image quality, but with paired comparison they extend their rating to other dimensions, including the image size. PC however has an important drawback: its cost and time consumption. To obtain scale value quality scores from PC data, two models exist: the Bradley-Terry model or the Thurstone-Mosteller model [110]. Both need a full PC matrix: each condition has to be compared to another. However, several efficient approaches have been developed in the literature to reduce the number of required comparisons [111], [16]. In [16] six video sequences were recorded. Each of these videos were captured at six inter-camera distances (10 cm to 60 cm). The 36 video sequences were then compared through paired comparison, and the Bradley-Terry scores of each condition were determined. Results show that the Bradley-Terry scores reveal quality fluctuations due to the different depth and comfort. The relation between inter-camera distance and QoE was found highly content dependent. In [112], 3D was compared to 2D using a PC approach on an auto-stereoscopic display. 3D was produced internally by the display based on a texture and a depth map. The texture was used at four different quality levels (three encodings and a reference). Results show that 3D was rejected in 70% of the cases and for the lowest quality rejected at 56%. However, the results may be influenced by the technology used at the time of the experiment and the quality of 3D rendering of the 3D display as mentioned by the authors [112].
Considering that PC provides an easy question to the test participant, it will be used to evaluate 3D and 2D video sequences, to show the quality improvement/decrease due to 3D. The research questions are the following:

**Listing 3.7: Research questions**

```
1. How much is 3D preferred (or less preferred) compared to 2D?
2. How does the preference of 3D over 2D evolves as a function of the image quality?
3. How much do the contents' characteristics affect the preference of 3D over 2D?
```

### *3.2.1 Definition of test conditions*

This subsection provides details on the definition of the conditions and test design to answer the previously stated research questions.

#### 3.2.1.1 Selection of Source sequences

The selection of the source contents (SRC) is based on three databases. All sequences were full HD stereoscopic videos; each view had a resolution of 1920x1080, with a frame rate of 25 images per second and were of 10s length. Seven SRCs come from a first database composed of 64 source reference signals (SRCs) which were evaluated in [113]. This database of SRCs will be further detailed in later stage of the thesis in the content characterization chapter.

The SRCs were used at the highest quality available, and contained various types of scenes. They were rated on three different scales: overall quality of experience, depth and visual comfort. The methodology used was Absolute Category Rating (ACR). Perceived depth was rated on a five-point scale with labels: “very high”, “high”, “medium”, “low” or “very low”. Using this general depth scale, the observers rated their general impression of the depth, which was thought to take into account both depth layout perceptions and depth quality. The comfort was evaluated in an absolute manner by asking subjects, if the 3D sequence is “much more”, “more”, “as”, “less”, “much less” - “comfortable than watching 2D video”. Based on this data, the seven SRCs were chosen to cover the entire range of depth ratings. To ensure the reproducibility of our results, it was decided to include five open video materials [114]. These sequences include “Tree Branches”, “Hall”, “Umbrella” and two new sequences designed to reach our depth effect requirements: one with low and high depth quantity, respectively, “Timelaps” and “Drone”. A third source of SRC was Blu-Ray disk where three other non-open sequences named “Alice” were added to the test. These last sequences were not available at the time of the previous test on content characterization described in [113]. Therefore the depth perception model [113] developed in the context of this thesis and which will be detailed in Chapter 5 was used to have a predicted value of the perceived depth in these sequences. The top scatter plot of Figure 3.24 shows how the selected sources cover the depth scale, based on subjective data [113]. The second scatter plot shows the results of the depth score estimated from the depth model described in [113]. The third scatter plot shows the available subjective data regarding visual discomfort. This data has been added, however, not used in this study for content selection, and is shown to let the reader have a view on the principal characteristics of the 3D sequences.

#### 3.2.1.2 Selection of coding conditions

Coding was performed using a Harmonic Electra 8000 H.264 encoder at constant bitrate. This encoder was the same as the one used in the previous study on video quality prediction models. Since it was planned to use a polarized display with horizontal interlacing, the 3D sequences were in the Top/Bottom frame compatible format, this choice limiting the loss of resolution. Each full-HD view was downscaled to half the vertical resolution using a Lanczos filter. No further interpolation was done for optimizing resolution, resulting in half a line of vertical parallax. Each sequence is then encoded to four different “quality levels” by using four different values of bitrate. The four bitrate values were individually defined for each source sequence, since each of them had different spatial and temporal complexity. In the previous section, it was revealed that VQM (ITU-R Rec. J.144) performs sufficiently well in estimating video quality of 3D sequences [94] (Pearson correlation of 0.89, and RMSE of 5.4). As a consequence, the procedure for determining the bitrate values corresponding to the appropriate “quality levels” is based on quality estimations obtained from the VQM general model. The finally used four adequate quality levels have been determined as 0.1, 0.2, 0.3, and 0.4 on the VQM scale. These values correspond to the quality score of the most complex sequence of the test, “TreeBranches”, encoded respectively at a quantization parameter QP of 26, 32, 38, and 44 using the reference H.264 encoder JM 18.2. The range of bitrates used in the test is illustrated in Figure 3.24 and noted for example 2DQ1, for 2D at quality level 1. Quality level 0 indicates the reference.

At the end of the selection process, the 15 SRCs were encoded at four individually chosen bitrates leading to VQM scores close to 0.1, 0.2, 0.3 and 0.4 as described above. These sequences are used in two versions: 2D and 3D. The 2D sequences were encoded at the same bitrate as their respective counterpart sequences in 3D. In addition, a 3D reference with no compression was added to the test. This results in $15 \cdot (2 \cdot 4 + 1) = 135$ video sequences to be evaluated.

### 3.2.2 *Evaluation of 3D QoE using paired comparison*

The global QoE of the video sequences was evaluated in a paired comparison experiment. 35 Observers participated in this test. The laboratory environment was in accordance with ITU-R Recommendation BT.500. The observers’ vision was screened in terms of acuity, color vision (Ishihara test), and stereo vision (Randot stereo test). For the test, two polarized 23” Hyundai displays (ViewSonic V3D231) with horizontal interlacing were used. The displays were

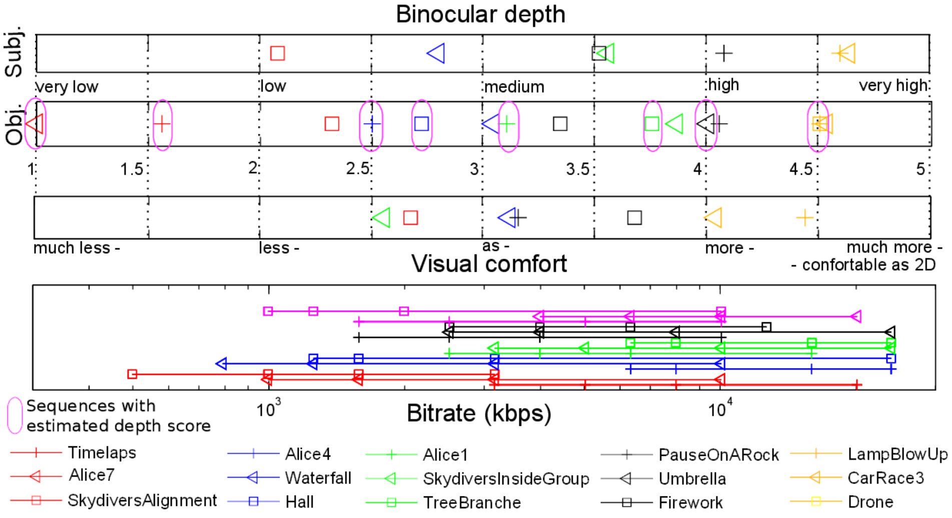


Figure 3.24: Depth, visual comfort, bitrate of source sequences

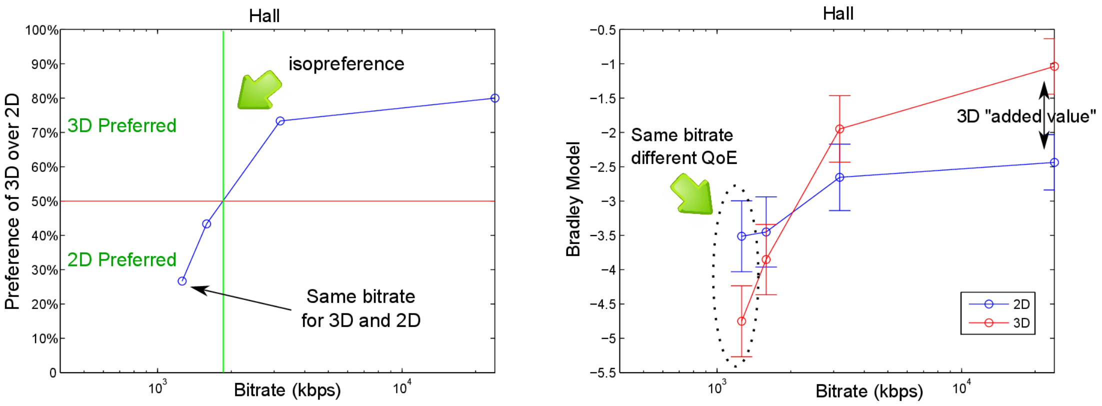
calibrated using a display calibration device (X-Rite i1 Pro) to make the rendering as similar as possible between the two displays. The observers were facing two distinct displays, and were instructed to give their preference between the two presentations they could see on the displays. Considering the number of possible presentations (3D or 2D, 4 quality levels in 2D and 3D each, and a 3D reference), a full PC matrix approach would have required $9 \times (9-1)/2 = 36$ comparisons per SRC, hence 540 comparisons for evaluating all video sequences. This high number of comparisons is impracticable for a subjective experiment [111]. For more efficient testing, the square design matrix was employed. Based on this approach, it was possible to reduce the number of comparisons to 18 comparisons per SRC, hence $15 \times 18 = 270$ comparisons in overall. The comparisons made in the test can be found in listing 1. The sequence pairs were randomized such that in case of comparison A vs. B, both orders A vs. B and B vs. A were seen by the observers. This avoids any dependency of the preference ratings on the display and possible default answers by observers (right vs. left). The test was split into two sessions of 45 min. The same observers participated twice, with a minimum time between the series of one week.

**Listing 3.8: List of sequence pairs compared by observers**

```
| 3DQ4  vs  3DQ0   |  2DQ3  vs  2DQ1  |  3DQ4  vs  2DQ4   |  2DQ3  vs  3DQ2  |
| 3DQ0  vs  2DQ4   |  2DQ1  vs  3DQ2  |  3DQ3  vs  3DQ1   |  3DQ4  vs  3DQ3  |
| 3DQ3  vs  2DQ2   |  3DQ4  vs  2DQ3  |  3DQ1  vs  2DQ2   |  3DQ3  vs  2DQ3  |
| 3DQ0  vs  3DQ1   |  3DQ0  vs  2DQ1  |  3DQ1  vs  2DQ1   |  2DQ4  vs  2DQ2  |
| 2DQ4  vs  3DQ2   |  2DQ4  vs  2DQ4
```

### *3.2.3 Preference of 3D over 2D and pictorial quality*

The main goal of the paired comparison test was to analyze how the preference of 3D over 2D would depend on the respective “pictorial quality”. As outlined in Section 3.2.1.2, pictorial quality was varied at four different bitrates, i.e. quality levels Q1-Q4. Figure 3.25a illustrates, for one SRC, how observers answered. It is visible that when the bitrate increases, the preference of the 3D presentation over the 2D version increases. This behavious is, however, different for one of the contents, “SkydiversInsideGroup”. This content was found to be less preferred when the quality increased. In Figure 3.24, it can be seen that this content is the least comfortable sequence of the database. The blurring added by coding may have contributed in such a way that this content was perceived as more comfortable when coding was stronger. This would be in agreement with [115], where binocular fusion was found to be dependent on the retinal disparities and spatial frequencies within images. In the paired comparison tests, some video sequences were always found to be preferred in 2D, namely “SkydiversAlignment”, “SkydiversInsideGroup”, and “Waterfall”. These sequences correspond to the least comfortable sequences of the test. In turn, one sequence was always preferred in 3D: “CarRace3”. Based on the test results, the bitrate at which 2D and 3D were equally preferred can be determined, as well as the respective VQM scores. In the following, these points are referred to in terms of *isopreference*. The VQM scores at isopreference has been estimated using linear regression between two known points in the 2D domain spanned by “preference percentage” and “VQM scores”. On average, it was found that the isopreference at the same bitrate between 2D and 3D is reached when the picture quality of the 3D sequence measured by VQM is at most equal to 0.24. The relation between the VQM scores at isopreference and the depth score rating was considered (subjective depth score when available, and objective if not, see Section 5.4). However, no simple relation can be found between these two factors, and other factors have to be taken into account. These other factors may include monocular depth cues such as blur from defocus, linear perspective, texture gradient, motion parallax and also visual discomfort.



(a) (b)

Figure 3.25: Illustration of preference results for one source content (Hall).

### *3.2.4 Quantitative analysis of the “3D added value”*

Thanks to the test design it was possible to use the Bradley-Terry (BT) model for the analysis. In Figure 3.25b, results of the model are depicted for one example SRC. The BT-scores provide the continuous perceptual scale which quantifies the difference between 2D and 3D QoE. It is then possible to evaluate the “added value of 3D” by measuring the difference between the BT-scores at conditions where the bitrate is the same. As only pairs for the same source content were evaluated in our test, the BT-scores cannot be used to compare preferences between contents. For example, it is not possible to compare the content “Alice1” in 3D at quality level 3 to “SkydiversInsideGroup” in 2D at quality level 2. This inter-content comparison was not targeted, and instead the goal was to determine 3D preference thresholds as a function of pictorial quality for different degrees of depth information. Making inter-content comparisons would have added individual judgments of the observer regarding his preference of one type of scene compared to another, which would have made the data noisy and hard to interpret. As a consequence, it is not possible to compare one BT-score from one SRC to another score from another SRC since there is an unknown offset between these two scores. However, since the scale remains the same between SRCs, it is then possible to compare inter-SRC differences of BT-score. Let the “3D added value” be the difference of BT-score between two similar coding conditions reflecting the

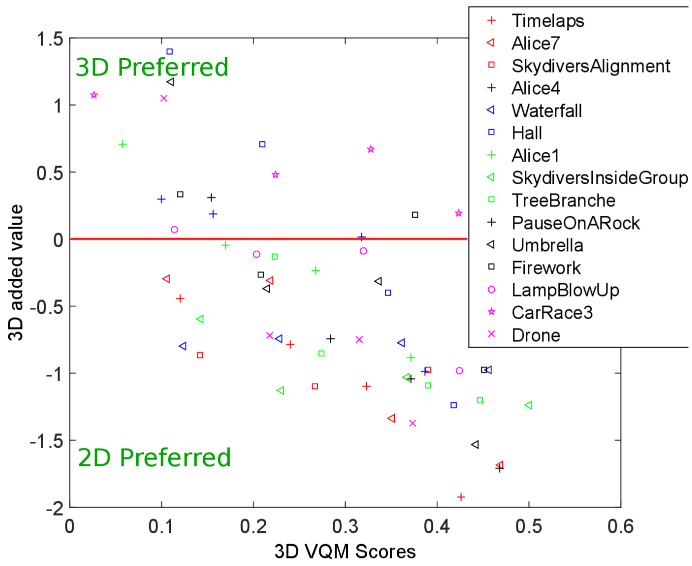


Figure 3.26: 3D added value as a function of VQM scores.

score fluctuation due to the presence of depth (see Figure 3.25b). Then, it can be seen that at least two factors are of influence on the "3D added value" (*3DAV*) scores, one covers the 3D characteristics of the video sequences including depth, comfort, naturalness, immersion, etc. And the other one covers the pictorial quality of the video. Figure 3.26 depicts the relation between the quality factor measured through VQM scores of the 3D video sequences and the *3DAV*. These two factors show a Pearson correlation of -0.65 and a Spearman correlation of -0.67. Using a N-Way Analysis of Variance (NANOVA) analyzing the *3DAV* based on the factors "QualityLevels" as defined in Section 2 and "DepthLevels" (grouping the SRCs in five classes of depth effect) shows that there is a strong influence of quality on the 3DAV $(F = 13.5, p < 0.001)$ and that there is also a significant influence of the "DepthLevel" on the *3DAV*

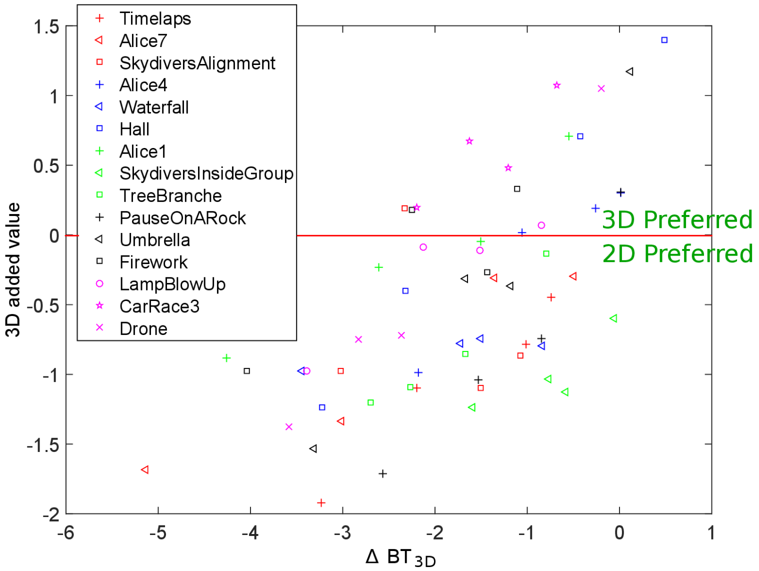


Figure 3.27: 3D added value as a function of $\Delta BT_{3D}$.

($F = 3.98, p = 0.0069$). Considering the rather small amount of data (four *3DAV* values per SRC), no significant influence can be observed on a per-content analysis. Let $BT_{3D}(k)$ be the BT-score of the condition *3DQk* as listed in listing 1. Then, the 3D-QoE impact due to coding can be obtained by: $\forall k \in [1,4], \Delta BT_{3D}(k) = BT_{3D}(k) - BT_{3D}(0)$. The $\Delta BT_{3D}(k)$ provides a subjective value of how coding affects 3DQoE. This includes loss in pictorial quality, loss in depth [11] and the impact on comfort. The relation between $\Delta BT_{3D}(k)$ and *3DAV* is depicted in Figure 3.27. For the latter scatter plot, it should be remembered that both 3DAV and $\Delta BT_{3D}(k)$ are expressed on the same scale. To compare content's specificities it is proposed to perform a regression of the 3D added value (*3DAV*) as a function of $\Delta BT_{3D}(k)$ (see Table 3.10).

The slope values ($\alpha$) between the two factors range from 0.02 to 0.76. The result for the overall data to a slope of 0.71. This shows that on average, a quality variation of $X$ will impact by $0.71 \cdot X$ the added value of 3D over 2D. However, there are high fluctuations due to content specificities (depth and comfort) which need to be studied further. The values of $\beta$ provides information on the added value of the content when available at the highest quality possible and then its suitability to be presented in 3D. Considering the high variation inter-content of $\alpha$ and $\beta$ a characterization of the scenes appears to be needed for the development of 3D-QoE models. This will be addressed further in chapter 5.2 of the thesis.

| $3DAV = \alpha \cdot \Delta BT_{3D} + \beta$ | | | | | |
|---|---|---|---|---|---|
| Content | $\alpha$ | $\beta$ | Content | $\alpha$ | $\beta$ |
| Timelaps | 0.54 | -0.10 | Alice7 | 0.33 | -0.07 |
| Sky.Alignment | 0.021 | -0.94 | Alice4 | 0.58 | 0.38 |
| Waterfall | 0.08 | -0.67 | Hall | 0.68 | 1.05 |
| Alice1 | 0.40 | 0.77 | Sky.InsideGroup | 0.38 | -0.72 |
| TreeBranche | 0.57 | 0.23 | PauseOnARock | 0.76 | 0.13 |
| Umbrella | 0.76 | 0.90 | Firework | 0.38 | 0.65 |
| LampBlowUp | 0.41 | 0.53 | CarRace3 | 0.51 | 1.33 |
| Drone | 0.71 | 1.15 | overall | 0.71 | 1.15 |

Table 3.10: Relationship between added value of 3D and difference of BT-score between coding and reference condition.

### *3.2.5 Relation with previous studies*

The relationship found between the "3D added value" and the 3D quality of the video is particularly interesting since this experiment found for a different context similar coefficients between these factors as in the work of Lambooij [9]. In Lambooij's work, test participants were asked to report their *Visual experience*, the 3D image quality, and the depth for video sequences having only blur and Gaussian noise distortions. This has resulted in the equation:

$$EC = a \cdot IQ + b \cdot D \tag{3.2}$$

With 0.74 and 0.26 respectively for $a$, and $b$ when the evaluation concept (EC) *naturalness* is modeled, and 0.82 and 0.18 were found for $a$, and $b$ when the *visual experience* is targeted. In this experiment, a coefficient of 0.71 was found for the 3D image quality factor when preference of 3D over 2D is modeled. However, the term $\beta$ from the study conducted in this thesis is most likely taking into account more factors than the depth since it includes all factors describing how appropriate the content is rendered in 3D.

### 3.2.6 Conclusion

3D QoE was evaluated by using paired comparison. This way, preference of 3D could be investigated as a function of picture quality. Results show that increasing picture quality increases the probability of preference of 3D over 2D. On average, a VQM score of 0.24 was found to be required to ensure preference of 3D over 2D. Bradley-Terry scores were estimated, and the "3D added value" was determined. The results show that, on average, there is a factor of 0.71 between variation of pictorial quality and "3D added value". However, there is lots of variation between contents, which need to be studied.

**Listing 3.9: Conclusion on the research questions**

```
1. & 3. In this experiment, it was possible to evaluate the preference of 3D over 2D.
    The ``appropriateness'' of contents to be represented in 3D was evaluated with the
    term β in Table 3.10
2. & 3. It was found that the preference of 3D over 2D increases when the pictorial
    quality increases.
```

## 3.3 Overall results

The last experiment presented has shown the relationship between preference of 3D over 2D as a function of image quality and content characteristics. A linear model has been found providing a good fit between these different factors (see eq. 3.3). The parameters $\alpha$, and $\beta$ are content dependent and relate to factors such as the perceived depth and visual comfort. This characterizes the contents across two axis:

- The added value of the content compared to 2D: $\beta$
- The criticality of coding for the content, e.g. the importance of texture quality to preserve the added value of 3D: $\alpha$

The quality factor was demonstrated in the first two studies to be closely related between 3D and 2D-presentation, and it was even shown that prediction algorithms designed to predict 2D quality works well for predicting 3D quality. Therefore there is a strong research interest for the characterization of the 3D video sequences properties. This will be the topic of the next chapters of the thesis.

$$Preference_{3Dvs2D} = \alpha \cdot 3DImageQuality + \beta \tag{3.3}$$

## 3.4 Key contributions

In this chapter, the evaluation of 3D QoE has been addressed and methodologies have been compared to study how the added value of 3D compared to 2D can be revealed. The main contributions are listed below:

- It was shown in the case of a study using absolute category rating for evaluating difficult concepts such, that test participants do not use the scales in the same manner. This is due to the fact that they may not all understand the scales equally for complex notions such as visual comfort.
- Results have shown that good prediction algorithm for 2D video quality can be applied to 3D quality after a linear transformation.
- Using Pair Comparison, it was possible to have a simple question for the test participants enabling them to fully understand their task. This enabled to draw the relationship between texture quality and preference of 3D over 2D.
- The preference of 3D over 2D increases when 3D image quality increases.

- A linear relationship between 3D image quality and preference of 3D over 2D can be drawn (eq. 3.3). This enables to characterize the contents itself across two axes: the added value provided by the content itself when not affected by coding, and how critical coding is for a given content.

# Chapter 4
# Subjective evaluation of depth

## 4.1 Introduction

The previous chapter showed that it is possible to reveal the added value of 3D videos compared to 2D using pairwise comparison. Even if the display condition were set to minimize visual discomfort issues [89], it is clear that the relation between preference of 3D over 2D and texture quality is highly content dependent and then depends on the depth properties of the content. Therefore, there is need to characterize the 3D properties of the 3D videos. For the characterization of the 2D properties such as the temporal and spatial complexity, recommendation have been defined by the ITU in ITU-T Recommendation P.910 [116]. It defines the spatial information (SI) and the temporal information (TI). But no indicators have been defined for 3D videos. In this section, the characterization of the 3D properties will be addressed using evaluation involving test participants. As previously described in the chapter on state of the art and its related section on depth perception, many factors are involved in the depth perception process both regarding the 2D and 3D properties of the contents. In this chapter different methods involving test participant will be presented for the evaluation of 3D-properties of 3D video sequences.

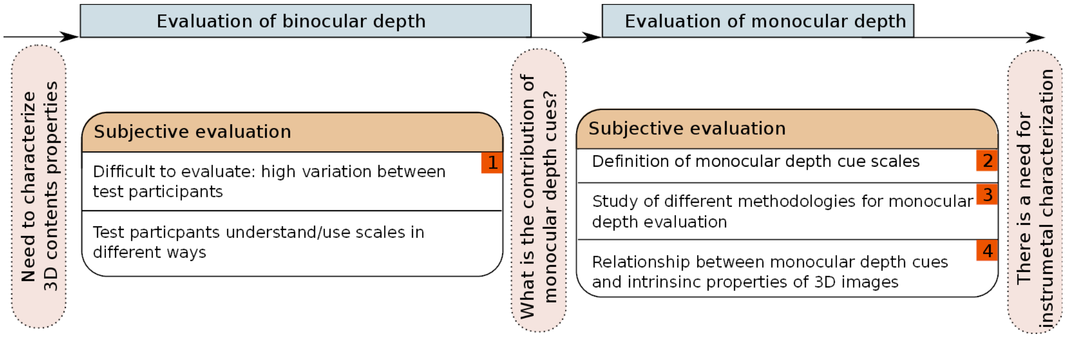

Evaluation of binocular depth
Evaluation of monocular depth
Need to characterize 3D contents properties
Subjective evaluation
1
Difficult to evaluate: high variation between test participants
Test particpants understand/use scales in different ways
What is the contribution of monocular depth cues?
Subjective evaluation
Definition of monocular depth cue scales
2
Study of different methodologies for monocular depth evaluation
3
4
Relationship between monocular depth cues and intrinsinc properties of 3D images
There is a need for instrumetal characterization

Figure 4.1: Structure of the studies described in the chapter.

## 4.2 Evaluation of depth cues

The evaluation of depth from monocular and binocular depth cues will be performed on natural images. The use of natural images makes the task particularly difficult since it becomes difficult to define the amount of each depth cue as usually done in psychophysical studies. It is then needed to evaluate how strong each considered depth cue is to enable the study of how they affect the overall perceived depth. This comes with a second difficulty: the fact that we

may omit to evaluate depth cues which are in the pictures and used by the test participants to evaluate the perceived depth. To limit this, all the different depth cues described by Cutting and Vishton [25] will be considered.
The evaluation of the monocular depth cues is challenging; most of the previously described methodologies in the state of the art section (See Section 2.3.1) were designed to evaluate the overall depth perception and not the individual depth cues. Indeed, these individual depth cues could be obtained thanks to the design of the experiment since there are mostly artificial scenes, and there was no need to evaluate these depth cue individually. The methodologies between the ones presented in the last subsection which enables the evaluation of the monocular depth cues are then the forced choice between a certain number of options and the evaluation on a numerical scale.
Unfortunately, to enable having a quantitative evaluation of a particular scale through a forced choice approach, a high number of comparison is required. Indeed either the Bradley-Terry model or the Thurstone-Mosteller model [110] requires a full pair comparison matrix and then $\frac{N\times(N-1)}{2}$ comparisons per scales with $N$ the number of stimuli. Optimized approaches are possible using the square design approach [117] and a well chosen square matrix [111]. However, considering that natural images were chosen and then the difficulty to define precisely the quantitative amount of monocular depth cues in the stimuli, a high number of images was selected: 200. Then even using the square design approach, the number of required comparison is still too high for being done in a subjective test.
Different approaches will be used to address the evaluation of the different depth cues and overall perceived depth. The research question addressed are the following, and were addressed though different experiments listed in Table 4.1.

**Listing 4.1: Overall research questions**

```
1. How 2D depth cues can be characterized in natural images?
2. How, by means of evaluation involving test participants, is it possible to measure
   the contribution of each individual depth cue?
3. How does the overall perceived depth relate to the different depth cues?
```

| Section | Depth cues | Type of content | Number of SRCs | Methodology | Published in |
|---|---|---|---|---|---|
| 4.4 | Binocular depth | 3D Videos | 64 | ACR | [113] |
| 4.5 | Monocular depth cues | 2D Images | 200 | ACR | [118] |
| 4.5.2 | Binocular depth | 3D Images | 200 | ACR | [118] |
| 4.6 | Monocular depth cues | 2D Images | 150 | Ranking | [119] |

Table 4.1: List of experiments conducted with addressed depth cues and addressed type of source sequences.

## 4.3 Definition of scales

A first important aspect is to provide a definition of the different scales, and define how the different depth cues should be rated by the test participants. A contribution described in this Section is the definition of the different scales which will be used by the test participants to evaluate the monocular depth cues of the 3D materials. Seven different depth cues were considered. Each of them was defined by a schema, several examples, and a definition. Each of the scales will be described in this section.

### *4.3.1 Perceived depth*

To evaluate the overall perceived depth, test participants were asked if the depth sensation was: “very high”, “high”, “medium”, “low” or “very low”. It was chosen not to ask participant to evaluate only binocular depth as natural images are considered, it was not possible for them to disambiguate the contribution of the monocular depth cues from the binocular cues. Therefore, only the overall depth sensation could be evaluated.

### 4.3.2 The linear perspective

Test participant were asked to evaluate the linear perspective (Figure 4.2) by taking into account if there are clear visible vanishing lines within the image and if these vanishing lines contributes to the perception of the different depth layers in the scenes. This depth cues is supposed to be stronger as clear vanishing lines are visible.

### 4.3.3 The relative size

Test participant were asked to evaluate the relative size (Figure 4.3) by considering if there are repeating objects in the scene which appears with difference size. They were insctruct not to use their knowledge about the size of the individual objects for the rating. The rate should have depended on the number of occurrence an object appears with different size. This depth cue is supposed to increase when objects are repeated several times at different sizes.

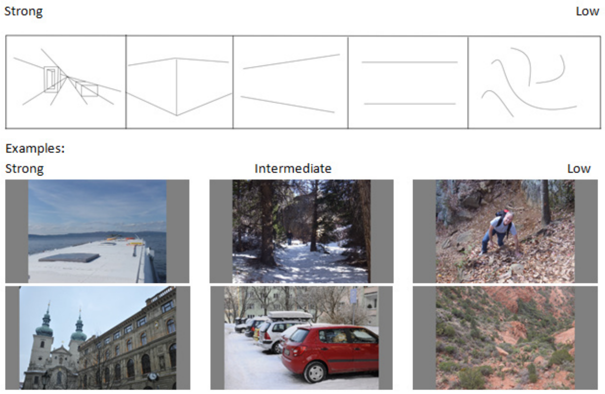


Figure 4.2: Linear perspective

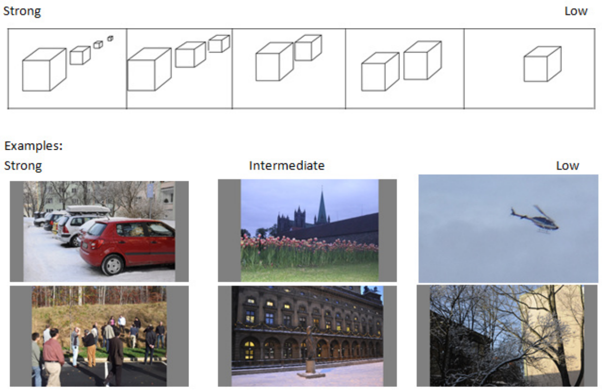


Figure 4.3: Relative size

### 4.3.4 The texture gradient

Test participant were asked to evaluate the texture gradient (Figure 4.4) based on the fact that there is a texture within the image (more generally can consider the repetition of patterns) which become finer when the distance to the camera increases. This depth cue is supposed to be stronger when there is a strong variation of the granularity of the texture or pattern.

### 4.3.5 *The interposition*

Test participant were asked to evaluate the interposition (Figure 4.5) based on the number of overlapping objects in the scenes. They were told that the overlap of one object over another provides the ability to order the position in depth of the objects and they should evaluate the interposition considering how the number of overlapping object helps to be aware of the absolute position in depth of the objects using all the interpositions. This depth cues is supposed to increase when there are a lot of objects overlapping at different absolute position in depth.

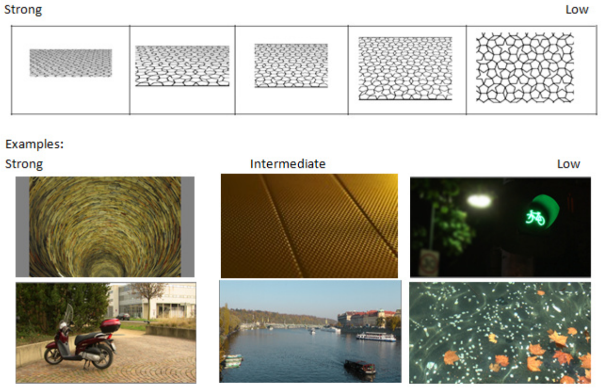


Figure 4.4: The texture gradient

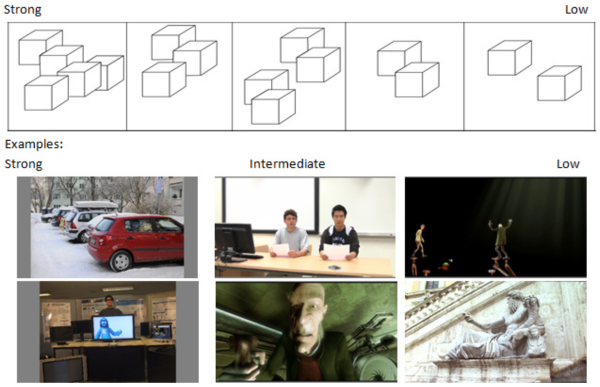


Figure 4.5: The interposition

### 4.3.6 *The light and shades*

Test participant were asked to evaluate the light and shades (Figure 4.6) based on the presence of a light source and the resulting shades which helps to apprehend the shape of the objects. This depth cue is supposed to be stronger when there is a light source which enables to see the real shape of the object which would have appeared flat otherwise.

### 4.3.7 *The areal perspective*

Test participant were asked to evaluate the areal perspective (Figure 4.7) based on the effect of the atmosphere in the image. For example, objects which are far away will have a color close to the color of the sky. This depth cue is supposed to be proportional to the presence of smooth transition of the color of the sky to the elements in the background which usually do not have this particular color of the sky.

### 4.3.8 *The defocus blur*

Test participants were asked to evaluate the defocus blur (Figure 4.8) based on the variation of the sharpness at different locations of the image explicating variation of the distance of the object to the focal point of the camera. This depth cue is supposed to be proportional to the variations between sharp and blurred area in the images.

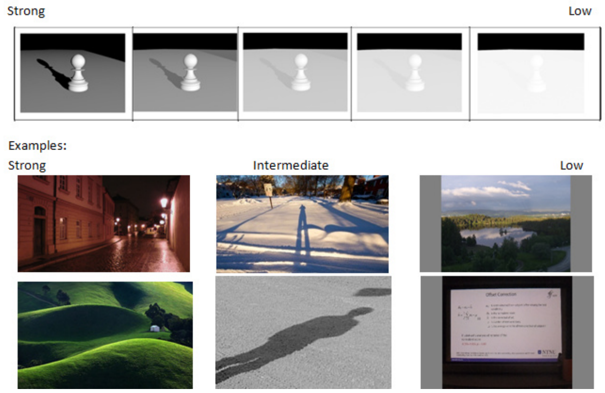


Figure 4.6: Light and shades

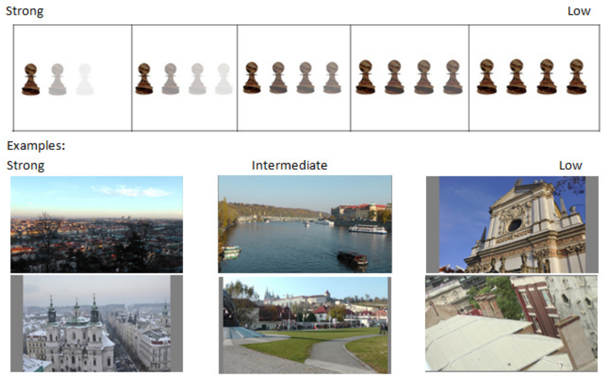


Figure 4.7: Areal perspective

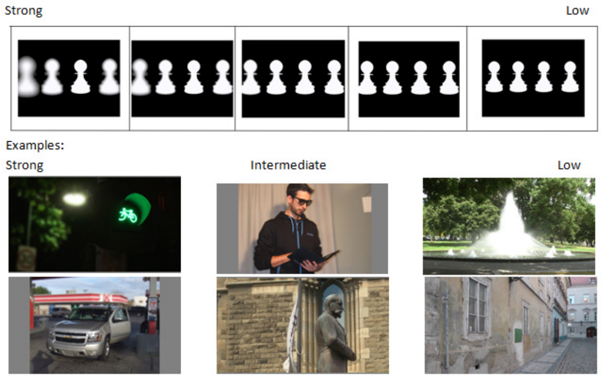


Figure 4.8: Defocus blur

## 4.4 Evaluation of perceived depth

In order to evaluate the monocular and binocular properties of the 3D video sequences and 3D images, different experiments have been conducted. These target to characterize the content properties, but also study how the different cues can be evaluated in natural images. The first experiment which will be described target to evaluate how the depth quantity can be evaluated in natural images and how test participants are consistent in the way they provide their ratings. The research questions are then the following:

**Listing 4.2: Research questions**

```
1. How depth in 3D images can be evaluated.
2. How do test participants agrees when asked to rate depth.
3. How test participants relate binocular depth, quality of experience and content
    characteristics.
```

### *4.4.1 Experiment*

Once the different scales were defined, different experiments were conducted to evaluate images and videos along the different scales. To evaluate the overall perceived depth, a database composed of 64 source reference signals (SRCs) has been designed [113]. A description of each source sequence can be found in Table 4.2. These SRCs were used at the highest quality available, and contained various types of scenes: indoor, outdoor, natural, or computer generated sequences, and containing slow or fast motion. The objective was to diversify at most the source material. All these sequences were full HD stereoscopic videos; each view had a resolution of 1920x1080, with a frame rate of 25 images per second. Each of the sequences was of 10s length. They were presented on a 23" LCD display (Alienware Optx, 120Hz, 1920x1080p). It was used in combination with the active shutter glasses from Nvidia (NVidia 3D vision system). The viewing distance was set to 3H, and the test lab environment was according to the ITU-R BT.500-12 recommendation [98]. Twenty four observers attended the experiment; their vision was checked, and it was assured that they passed the color blindness test (Ishihara test) and the depth perception test (Randot stereo test). Subsequently, they pass all the vision tests, the observers were trained using five sequences with different values of image quality, depth quality and visual discomfort. During the training phase the observers had the opportunity to ask questions. After the training had finished, the observers were asked to rate the 64 sequences on three different scales: overall quality of experience, depth and visual comfort. The methodology used was Absolute Category Rating (ACR). QoE was rated on the standardized five grade scale: "Excellent", "Good", "Fair", "Poor", "Bad". Perceived depth was rated on a five-point scale with labels: "very high", "high", "medium", "low" or "very low". Using this general depth scale, the observers have rated their general impression about the depth, which takes into account both depth layout perception and depth quality. The comfort was evaluated by asking subjects if the 3D sequence is "much more", "more", "as", "less", "much less" - "comfortable than watching 2D video". The test subjects were not presented with 2D versions of the video sequences, therefore they had to compare the 3D comfort with their internal references of 2D sequences. One test run took approximately 50 minutes, including the training session and a 3 minutes break in the middle of the test.

## *4.4.2 Analysis of results*

### 4.4.2.1 Agreement between observers

The coherence of individual ratings of each observer with those of the other observers was checked by following the $\beta_2$ test as described in section 2.3.2 from ITU-R BT.500 [98]. The screening was done for each of the three scales individually. Observers could be kept for a specific scale but rejected for another. This was motivated by the fact that observers may have misunderstood one scale, but may still correctly evaluate for the other scales. After screening, four observers of the 24 were rejected on each scale: two observers showed strong variation compared to the rest of the group on the quality and depth scales (according to the $\beta_2$ test), two on the comfort and quality scales, one on the depth and comfort scales, one on only the comfort scale and one on only the depth scale. None of the subjects showed inconsistent behavior for all three scales.

To further study the agreement between observers the correlation between test participants compared to each other was considered and are depicted in Figure 4.9. It appears that the correlation between every participants compared to each other is rather low. This appears to be due to the difficulty of test participants to evaluate the different factors on the proposed categorial scales. To analyze the differences of agreement between the ratings on the different scales, a KruskalWallis one-way analysis of variance is applied to compare the Spearman correlation between test participants as a function of the scale under evaluation, and a significant difference between the agreement of test participants on the three scales can be observed (Chi-sq=10.54, $p<0.01$). A Fleiss Kappa test was applied to compare the different agreements between observers when rating on the different scales and shows a higher agreement of the test participants on the evaluation of the binocular depth (0.084) followed by the evaluation of visual comfort (0.081), and the evaluation of quality (0.08). Although all Kappa values are low, an interpretation of the lower agreement for the quality scale may be interpreted by the fact that the task may have been hard to do for the test participants since the videos were presented at the highest quality possible, and therefore the evaluation concept of quality may have been unclear and resulted in more variation between observers.

| Sequence | Description | Sequence | Description |
|---|---|---|---|
| Alignment | NAT, skydivers building a formation together, low texture | BalloonDrop | NAT, balloon of water hit by a dart, closeup |
| Bike | NAT, cyclers, slow motion, lots of linear perspective | BloomSnail | NAT, closeup on flowers and snail, high depth effect |
| Building | CG, circular movement around towers | CarEngine | CG, car engine, many moving objects, high disparities |
| CarMesh | CG, car mesh rotating, low spatial complexity | CarNight | NAT, dark, many scene cuts (5), fire blast popping out |
| CarPresent | CG, circular movement around car | CarRace1 | NAT, race, rain, fast motion, several scene cuts (7 in 10s) |
| CarRace2 | NAT, race car, fast motion, several scene cuts (7 in 10s) | CarRace3 | NAT, race car, dust slowly flying towards the camera |
| Castle | NAT, highly textured, temporal depth effect changes | CristalCell | CG, many particles, different objects in depth |
| FarClose | NAT, skydivers, complex motion, increasing depth effect | FightSkull | CG, fast motion, low spatial complexity, high depth effect |
| FightText | CG, slow motion, objects popping out | Figure1 | NAT, skydivers, complex and circular motion, closeup |
| Figure2 | NAT, skydivers, complex motion, closeup, persons in circle | Figure3 | NAT, skydivers, complex motion, closeup group persons |
| Fireworks | NAT, dark, lots of particles, good depth effect | FlowerBloom | NAT, closeup on flowers, high depth effect |
| FlowerDrop | NAT, closeup on flowers and raindrop | Grapefruit | NAT, trees, highly textured, pan motion, high depth effect |
| Helico1 | NAT, low texture, circular motion, low depth effect | Helico2 | NAT, medium texture, circular motion, low depth effect |
| HeliText | NAT, medium textured, text popping out of the screen | Hiker | NAT, highly textured, person walking in depth |
| Hiker2 | NAT, highly textured, slow motion, closeup on persons | InsideBoat | CG, indoor, walk through the interior of a ship cabin |
| IntoGroup | NAT, pan motion, colorful, lots of objects in depth | Juggler | NAT, high spatial complexity, closeup on juggler |
| JumpPlane | NAT, skydivers, fast motion in depth (far from camera) | JumpPlane2 | NAT, skydivers, fast motion in depth |
| LampFlower | NAT, light bulp blowing up, flower blooming, closeup | Landing | NAT, fast motion, high texture, depth effect increasing |
| Landscape1 | NAT, depth effect limited to one region of the image | Landscape2 | NAT, depth effect limited to one region of the image |
| MapCaptain | CG, captain, map, slow motion, low spatial complexity | NightBoat | NAT, dark, low texture, camera moving around boat |
| Paddock | NAT, race setup, high spatial complexity, lots of objects | PauseRock | NAT, bright, closeup on persons sitting |
| PedesStreet | NAT, street, linear perspective, lots of motion in depth | PlantGrass | NAT, closeup on plant growing, grasshopper |
| River | NAT, slow motion, medium texture, boats moving | SkyLand | NAT, skydivers, high texture, person moving closer, closeup |
| SpiderBee | NAT, slow motion, closeup on spider eating a bee | SpiderFly | NAT, closeup on fly, spider and caterpillar |
| SpinCar | CG, car spinning, half of the car in front of screen | StartGrid | NAT, separate windows showing different race scenarios |
| StatueBush | NAT, closeup on statue with moving flag | StreamCar1 | NAT, high spatial complexity, car moving in depth |
| StreamCar2 | NAT, high spatial complexity, closeup on a car | StrTrain1 | NAT, train coming in, motion in depth, high textures |
| StrTrain2 | NAT, train coming in, motion in depth, many objects | SwordFight | NAT, sword fight, movement limited to one area of image |
| Terrace | NAT, persons chatting, camera moving backward | TextPodium | NAT, rain, fast motion, champaign and text popping out |
| TrainBoat | NAT, train and boat, fast motion in depth, medium texture | Violonist | NAT, closeup on violinist and her instrument |
| WalkerNat | NAT, persons walking between trees | Waterfall | NAT, closeup on water falling, highly textured |
| WineCellar | NAT, low spatial complexity, indoor, closeup on persons | WineFire | NAT, closeup on a glass and fire, complex motion |

Table 4.2: Description of the source sequences. CG: Computer generated, NAT: Natural scene

#### 4.4.2.2 Correlation between scales

The results show a high correlation between the different scales (Figure 4.10). The three scales are closely related: a Pearson correlation of 0.74 is observed between QoE and depth, 0.97 between QoE and visual comfort, and 0.71 between depth and visual comfort. The very high correlation between QoE and visual comfort could be explained as follows:

- It is worth pointing out that the video do not contain coding artifacts, so it is likely that people have rated the QoE of the sequences according to the sources of disturbance they perceived: the visual discomfort. Indeed, in presence of high disparity values as it may happen for sequences with a lot of depth, it may become more difficult for the

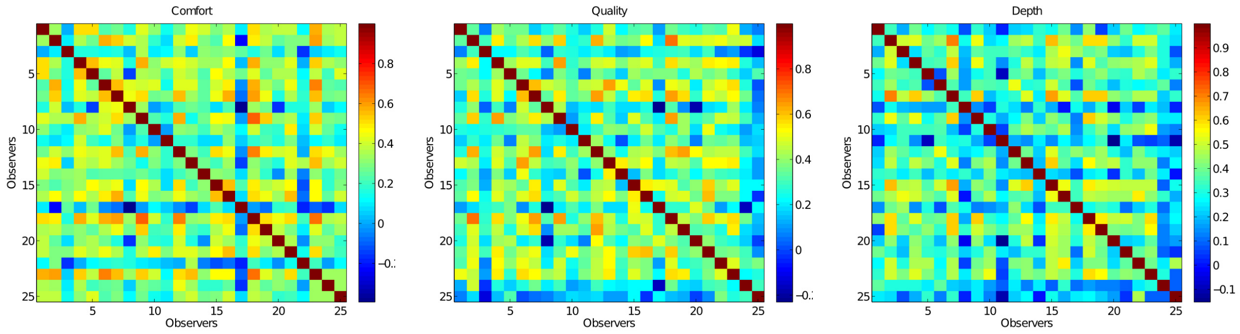


Figure 4.9: Spearman correlation between test participants

observers to fuse the stereoscopic views [120] [121]. This results in seeing duplicate image portions in distinct areas of the videos and is likely to be transferred to the quality rating.

- Another alternative explanation is that observers did not really understand the visual discomfort scale. This aspect has been addressed previously in Section 3.1.2.2 and by Engelke [104]. It has been observed that different classes of observers exist that differ in their understanding and thus use of the comfort scale. In this study, it is possible that observers have decided to use the comfort scale based on their QoE ratings.

It may be observed that there is a high variance between the source sequences in the here considered degradation-free case. The observed difference may be due to the content properties related to shooting and display conditions.
The lower correlation between depth and visual discomfort shows that there is no straightforward link between binocular depth and visual comfort, although both of them depend on retinal disparities.

### 4.4.3 Conclusion

The first conclusion of this study is that binocular depth is not necessarily easy to evaluate by test participants. Results have shown, that asking participants to evaluate depth results in lots of variation between participants, and therefore the evaluation on a categorial scale may not be the easiest approach for participants to evaluate difficult concepts such as depth or visual comfort.
This result was also confirmed by Engelke [104] who also found that observers use the scales in different manners showing the fact that they can have different interpretation of the evaluation concepts under study.
Finally, the last result of this study is a high correlation between the scales. The QoE scores are found to be highly related to the comfort scale. This may be due to the fact that no degradation were applied to the videos and discomfort appeared to be the main source of decrease of QoE. The correlation between visual comfort and depth was found much lower showing that these two factors are related but other factors are involved.

**Listing 4.3: Conclusion on the research questions**

```
1. & 2. The evaluation of 3D factors such as the depth is challenging, asking test
    participants to rate it on a scale results in variation amongst participants.
3. The result have shown a strong correlation between the different scales under
    evaluation. QoE and visual discomfort have shown a Pearson correlation as high as
    0.97, depth and visual discomfort was lower correlated with a Pearson correlation of
     0.71, and QoE and comfort show a Pearson correlation of 0.74.
```

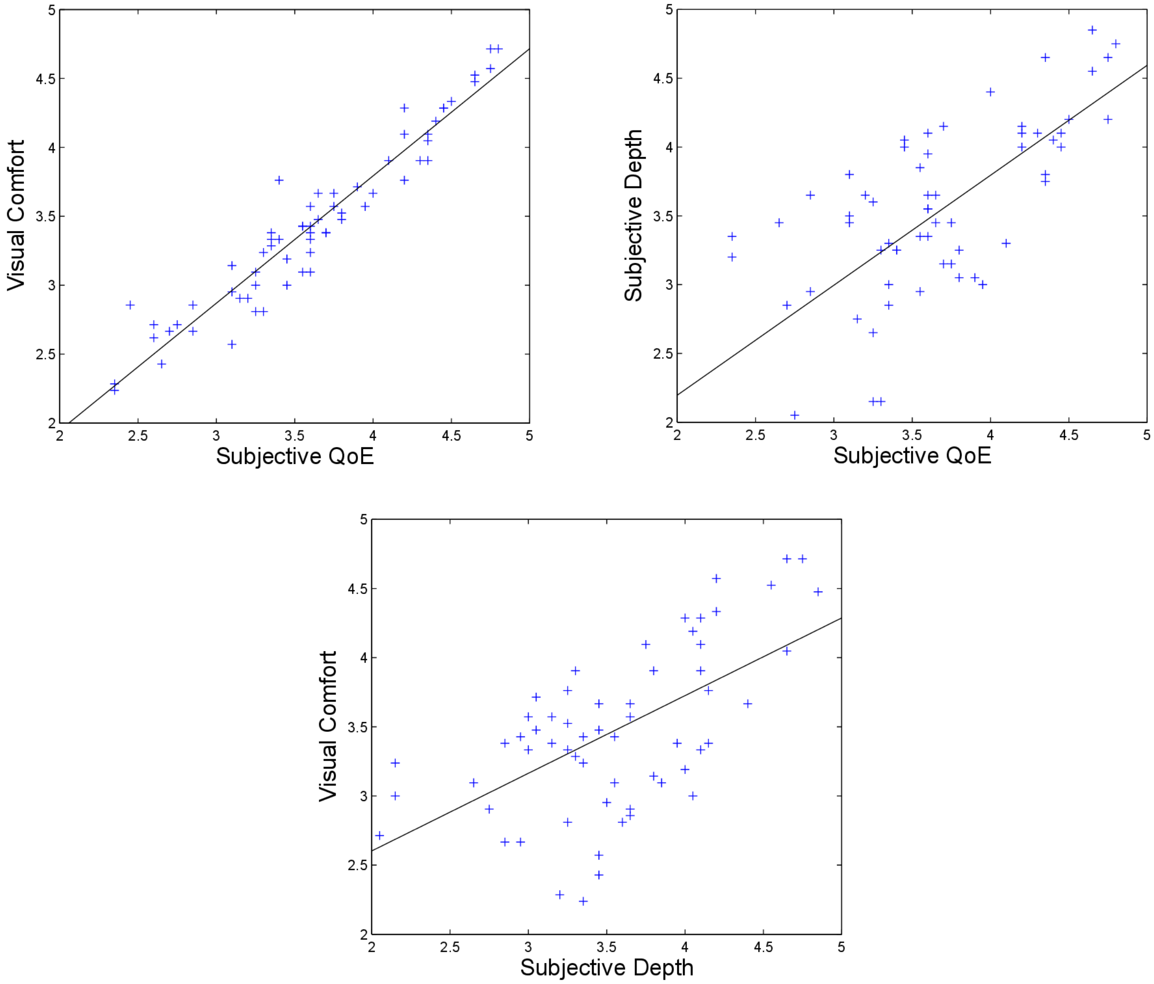


Figure 4.10: Scatterplots with regression lines showing the relation between the different evaluated scales

## 4.5 Evaluation of monocular depth cues

Previously, the overall perceived depth was evaluated. This section will address the evaluation of the different depth cues which contribute to the overall depth perception. There are different goals to this task: the characterization of the 3D videos properties taking into account perceptual factors, the study of how different depth cues can contributes to the overall depth percetion, and the relation between the different depth cues. In this first section, it will be addressed to which extent Absolute Category Rating (ACR) can enable to evaluate monocular depth cues in natural images. The research questions of this first study are:

**Listing 4.4: Research questions**

```
1. How image selection can be performed in order to study the relationship between depth
     cues and perceived depth in natural images.
2. Study the relationship between depth cues.
3. Study how depth cues relate with the overall depth perception.
```

### *4.5.1 Image selection*

To perform the image selection process, 409 images with a large variety of content were evaluated by two expert observers on the 7 different scales as described in Section 4.3 on a five grade category scales depicted in the Figures 4.2-4.8 and an evaluation of the binocular depth quantity on a five grade scale. The images were taken from different open source image and video database [122, 123, 124, 125, 126, 127, 114, 128] and images extracted from newly shot video sequences using a Panasonic AG-3DA1E twin-lens Camera, and new images shot with a Fujifilm FinePix Real 3D. After being evaluated on the different scales, it was decided to select four depth cues which will be studied as independently as possible: the linear perspective, the relative size, the texture gradient and the defocus blur. For each of these four depth cues the images were selected such that the score for the six other monocular depth cues was as small as possible and the values of the monocular and binocular depth cues range uniformly distributed from 1 to 5. This results in a matrix of five by five images as shown in Figure 4.11. Such matrix is then defined for the four previously mentioned depth cues. This results in selecting 100 different images. Considering that the pre-test phase used for the selection of images is only made with two expert observers, it was decided to add a repetition of each combination of monocular and binocular depth cue to increase the robustness of the image selection process and the likelihood to get the expected combination of monocular and binocular depth cues. This finally results in 200 images.

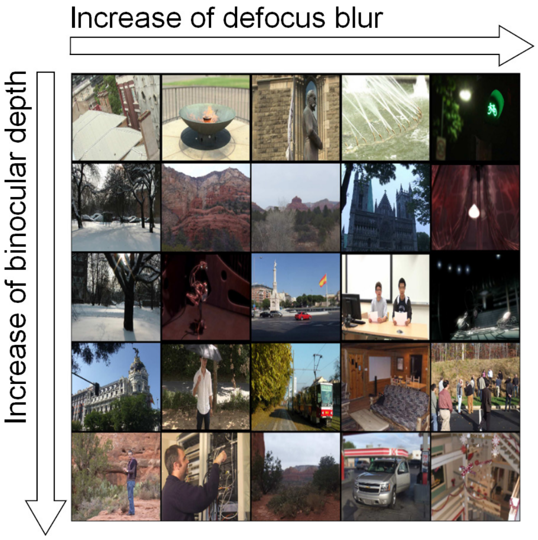


Figure 4.11: Example of matrix defocus blur / binocular depth

### *4.5.2 Evaluation of binocular depth*

The evaluation of the binocular depth quantity was performed in a first subjective experiment. The 200 still images were used at the highest quality available and did not have any visible coding artefact. Multiple open image and video databases, and new recorded content from a video camera or compact cameras were used. The images have different formats: The images coming from the camera having an aspect ratio of 4:3 were downscaled from the resolution 3648x2736 to 1440x1080 and were inserted in a uniform gray frame of 1920x1080. The images coming from the database 3DIQA [122, 123] where slightly smaller than 1920x1080 and were then centered in a uniform gray frame of 1920x1080. The other images having natively the resolution of 1920x1080 were kept in their original format. The images were presented on a 3D stereoscopic display: Samsung UE46F6500, 46” smart TV with active glasses and a native resolution of 1920x1080. The viewing distance was set to 3H, and the test lab environment was according to the

ITU-R BT.500-12 recommendation [98]. Twenty observers attended the experiment; their vision was checked, and it was assured that they passed the color blindness test (Ishihara test) and the depth perception test (Randot stereo test). The observers were trained using five different images with different amount of depth quantity. During the training phase the observers had the opportunity to ask questions. After the training had finished, the observers were asked to rate the 200 images on the amount of perceived depth. The methodology used was Absolute Category Rating (ACR). Perceived depth was rated on a discrete eleven grade scale from 0 to 10 with the labels “very high”, “high”, “medium”, “low” or “very low” - perceived depth respectively at the position 9, 7, 5, 3, 1 on the scale.

### 4.5.3 Evaluation of monocular depth

The evaluation of the monocular depth cues was performed using the Absolute Category Rating method on a five point scales to evaluate the seven scales corresponding to each individual depth cue (linear perspective, texture gradient, interposition, relative size, light and shade, areal perspective, defocus blur). The test participant were given the instructions described in Subsection 4.3, this includes the text describing the depth cues, the pictograms showing the different amount of depth cues and the examples images. The 200 images had the same HD resolution, as described in subsection 4.5.2 and were displayed on an 9.6 inches iPad 4 with a native resolution of 2048x1536. The images were presented on the top of the interface and were as large as possible and the different scales were represented as pictograms below the image under evaluation. Figure 4.12 shows the test interface of the application. Test participants evaluated each individual depth cue by selecting the pictograms. The application switched then automatically to the next scale. Test participants were able to edit a previous rating. Once all the seven scales were evaluated, a button appeared allowing the test participant to switch to the next image. For this test, no time constraint was given. The instructions were printed allowing the test participant to refer to them anytime they wanted. For this test, 8 experts in video or audio quality assessment participated to the test. The devices were given to the test participant and the test was performed in a non controlled environment. On average, the test requires 3 to 4 hours and was completed in several sessions within a week at the convenience of the test participants.

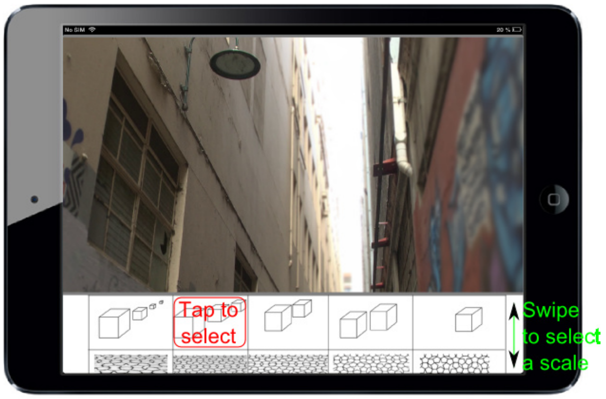


Figure 4.12: Subjective test interface

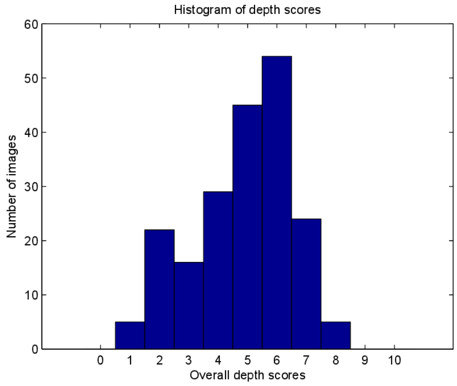


Figure 4.13: Histogram of depth scores from the 3D viewing session

### *4.5.4 Result*

#### 4.5.4.1 Analysis of vote distribution

First, regarding the scores from the 3D viewing session for the 200 images, a histogram built by categorizing the mean score over the observers per image can be found in Figure 4.13. The purpose of this histogram was to have a view on the distribution of the scores. There are 11 bins in this histogram to correspond to the 11 different categories that the observer had. 50% of the images were rated with a score higher or equal than 5. A Jarque-Bera test shows that the distribution of the subjective scores is not normal at a 95% confidence. Similarly, the histogram of the monocular depth cues scores is depicted in Figure 4.14. The selection of the images was done such that the distributions of the monocular depth cues linear perspective, relative size, texture gradient and defocus blur were meant to cover the entire range. The minimum score for each depth was not frequently used, and the average scores for each depth cue span from 2 to 5. For a particular depth cue, it was expected by design that 160 images would only show a small amount of this particular depth cues, corresponding to the first bin, and that 10 images would be voted with slight, medium, advanced and strong depth cue, respectively, filling the bins 2,3,4 and 5 with 10 samples each. This would have been one of the conditions of a good separation between the depth cues: one depth cue which changes of value in a controlled manner from 1 to 5 while all the other depth cues are kept to a minimum value. This kind of result has been achieved for the defocus blur. The areal perspective depth cue shows also a similar pattern. The texture gradient follows this rule to a lower extent. Regarding the linear perspective and the relative size, the distribution is more uniform, this shows that the image selection did not succeed to decorrelate the increase of other depth cues and the increase of relative size or linear perspective. For example the selection of images increasing the amount of texture gradient may have resulted in images having higher amount of relative size. Figure 4.15 depicts the Spearman correlation between the monocular depth cue scores. The correlation values are low, and indicate that the depth cues scores have only little relation between each others. The correlation values however support the discussion about the unexpected distribution of the linear perspective and relative size scores which have a higher correlation between each other showing that these two and the texture gradient and interposition may have shared some images even though the correlation values are too low to be conclusive.

#### 4.5.4.2 Relation between monocular depth cues and overall depth scores

One of the objectives of the study is to evaluate if in the context of the use of natural images it is possible to show the effect of monocular depth cues on the overall depth score values. One strong limitation of the study is due to the use of natural images which results in the absence of ground truth for the binocular depth cue. Indeed, the subjective data coming from the test in stereo mode described in Subsection 4.5.2 provides the results of the depth score rating resulting from the combination of monocular depth cues and binocular depth cues. The test from the subjective test described in Subsection 4.5.3 only provides monocular depth cue scores. The binocular depth cues themselves could not be controlled as usually done in psychophysics studies since natural image content was used. It is then only possible to use statistics about depth maps and content characteristics as described in previous studies [113] to retrieve information about the binocular depth cues.

To analyse the contribution of the monocular depth cue, for each depth cue, two categories are created: one with a low value of the particular depth cue and one with a high value of a depth cue. Let $\Omega$ be the set of all images. $DC_{c,high}$ is the set of images such that the depth cue $c$ is *high*. And $\forall I \in \Omega, DC_c(I)$ is the value of the depth cue $c$ for the image $I$.

$$DC_{c,high} = \{I \in \Omega | DC_c(I) > 3\} \tag{4.1}$$

$$DC_{c,low} = \{I \in \Omega | DC_c(I) \leq 3\} \tag{4.2}$$

To study the effect of a depth cue $c$ on the overall depth, the differences between the overall depth scores of the sets of images $IS1(c)$ and $IS2(c)$ are studied. $IS1(c)$ is the set of images where the depth cue $c$ is high and all the other depth cues are low. $IS2$ is the set of images where the depth cues are low.

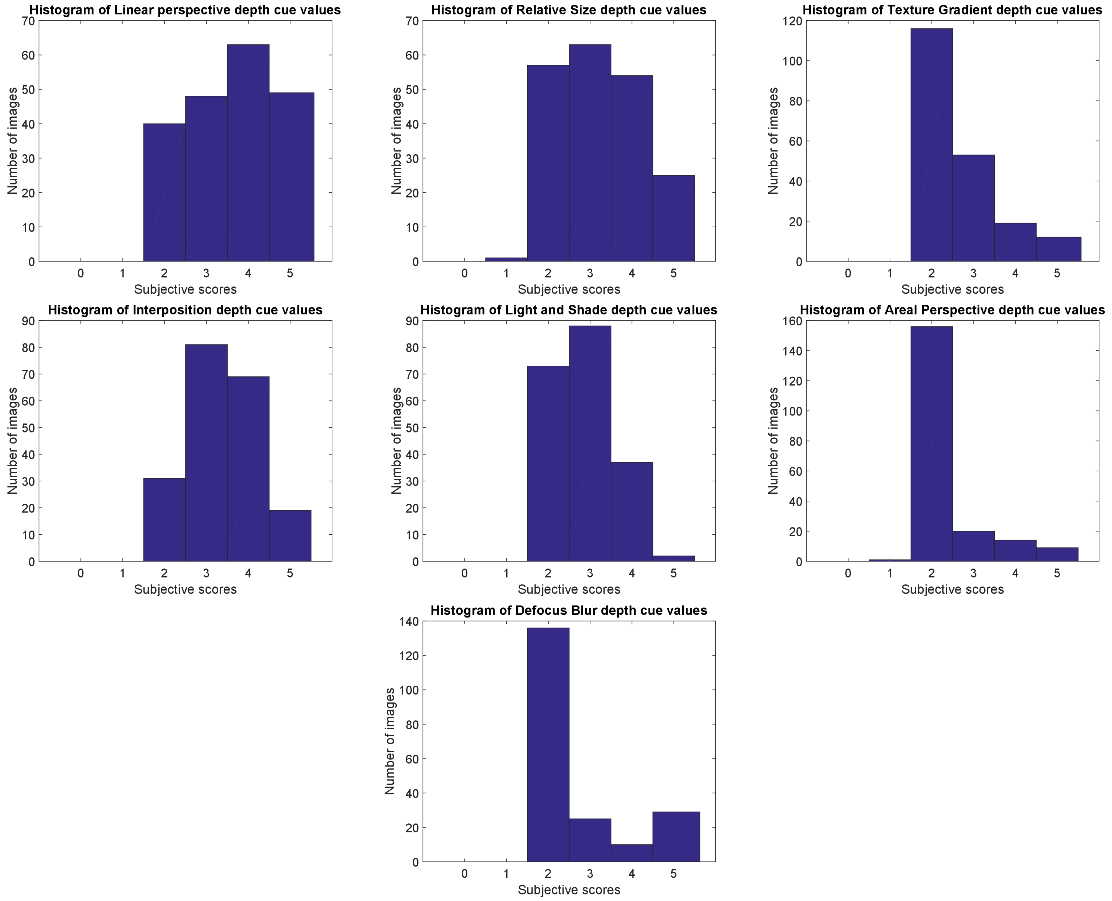


Figure 4.14: Histogram of monocular depth cues scores for the different depth cues. This represents the number of images which have a specific depth quantity for each depth cues.

$$IS1(c) = DC_{c,high} \bigcap \{ \bigcup_{i\in\{'LP','RS','TG','I','LS','AP','DB'\}\setminus c} DC_{i,low} \} \tag{4.3}$$

$$IS2 = \bigcup_{i\in\{'LP','RS','TG','I','LS','AP','DB'\}} DC_{i,low} \tag{4.4}$$

Unfortunatly, an ANOVA cannot be performed for each depth cue $c$ in order to compare the depth scores between the two set of images $IS1(c)$ and $IS2$ because the residual of the linear model using only one factor does not fulfill the normality requirements. As shown in Figure 4.15, the correlation between the scales is low, a PCA applied to the data confirms this result and shows that the explained variance increases linearly with the number of support vectors. It is then difficult to decrease the dimensionality. A linear model using all the different variables with no interaction term is then suggested as discussed in the state-of-the-art chapter weak models are a popular approach for depth cues fusion. Using a Jarque-Bera test, it is possible to confirm that the residual error of such model is normal. Table 4.3 lists the coefficients of the model. It is then possible to apply an N-Way ANOVA to explain the overall depth scores as a function of the monocular depth cue scores. Only the interposition (F=20.75,$p < 0.01$) and the defocus blur, which is on the borderline was found to have a significant effect (F=3.93,p=0.049). Followed by the texture gradient (F=2.24,p=0.13) and “light and shade” (F=1.6,p = 0.20) which were not significant on a 95% confidence.

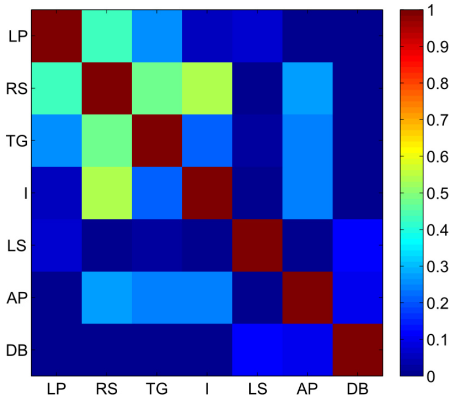


Figure 4.15: Spearman correlation between monocular depth cue scales; LP: Linear perspective, RS: relative size, TG: texture gradient, I: Interposition, LS: light and shade, AP: areal perspective, DB: defocus blur

| | LP | RS | TG | I | LS | AP | DB |
|---|---|---|---|---|---|---|---|
| LP | 100 | 42 | 26 | 5 | 7 | -13 | -17 |
| RS | 42 | 100 | 43 | 54 | 1 | 27 | -18 |
| TG | 26 | 48 | 100 | 22 | 2 | 24 | -12 |
| I | 5 | 54 | 22 | 100 | 0 | 24 | -1 |
| LS | 7 | 1 | 2 | 0 | 100 | -8 | 11 |
| AP | -13 | 27 | 24 | 24 | -8 | 100 | 10 |
| DB | -17 | -18 | -12 | -1 | 11 | 10 | 100 |

Figure 4.16: Spearman correlation between monocular depth cues; LP: Linear perspective, RS: relative size, TG: texture gradient, I: Interposition, LS: light and shade, AP: areal perspective, DB: defocus blur

$model = a \times LP + b \times RS + c \times TG + d \times I + e \times LS + f \times AP + g \times DB$

| | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| | 2.00 | 0.44 | 0.97 | 2.86 | 1.41 | 1.48 | 1.21 |
| F | 0.03 | 0.67 | 2.24 | 20.75 | 1.6 | 0.05 | 3.93 |
| p | 0.85 | 0.41 | 0.13 | $< 0.01$ | 0.20 | 0.83 | 0.049 |

Table 4.3: Linear model between depth cues. And (F,p) values of the N-Way ANOVA.

### *4.5.5 Limitations*

As mentioned previously, one limitation of the study is dependency on the binocular depth cues which can hardly be evaluated individually in natural images. During the design of the experiments, statistical analysis of the depth map characteristics were performed [113] to have a high variety of the content's stereoscopic properties. A further aggravation of this issue is also the limitation of the study to the particular instantiation of the problem, even though it was targeted by design to cover as much as possible the different monocular depth cue scales. These are limitations which could not be avoided in the targeted challenge due to the choice of natural images.

### *4.5.6 Conclusion*

The objective of this study was to reproduce studies performed in the field of psychophysics but in the particular case of the evaluation of monocular and binocular depth cues in natural images. Methodology questions have been addressed to tackle this challenge. A definition of different scales for the evaluation of monocular depth cues in images was proposed. Various analysis were performed to check the influence of monocular depth cues on the overall depth scores and to see if the methodology used could perform such task. Statistical differences of overall depth ratings between images showing low and high value of monocular depth cues could be seen for the particular case of the interposition and defocus blur depth cues, but not for the other depth cues. The image database (Depth Cue 3D Images, DC3DImg) including subjective scores has be made available, these can be used for example to study depth in 3D images, but can also be used for the investigation of other aspects such as the effect of coding on depth perception, the acceptance of

3D, the relation between monocular and binocular depth cues and depth quality issues, visual comfort and any other topics related to 3D quality of experience.

**Listing 4.5: Conclusion on the research questions**

```
1. The proposed image selection procedure enabled to select images with the expected
    properties. Subjective results were found consistent with the design of the
    experiment.
2. The Pearson correlation between ratings for the different depth cues was low.
    Relative size and Interposition were found to be better correlated.
3. It was not possible to draw a weak fusion model from the data. A classification of
    the scores in two categories: weak and strong only show an effect of the category ``
    low'' or ``high'' on the overall depth ratings for cues ``defocus blur'' and ``
    interposition''.
```

## 4.6 Alternative methodology: ranking

As described in the previous section asking directly test participants to evaluate the different depth cues may not be the most appropriate approach since the task is particularly demanding and scales difficult to evaluate. In order to tackle these issues, the previously described experiment have involved expert observers in order to reach higher accuracy, since they would show higher dedication to the task and are able to understand the complex notions to evaluate. However, even for expert observer the task is not easy to perform. Therefore an alternative methodology have been proposed in order to improve the understanding of the test participants of the task. The proposed method is based on ranking. The research questions are the following:

**Listing 4.6: Research questions**

```
1. How this methodology compare with traditional absolute category rating (ACR)
2. Does seeing the result of another test participant enable a better understanding of
    the test participants of the task they have to perform.
3. Does reordering the result of another test participant enable to improve the
    consistency between test participants.
```

### *4.6.1 Description of the proposed methodology*

#### 4.6.1.1 Description of the task

The new methodology proposed to tackle the previously described issues is depicted in Figure 4.17. The proposed task is divided into two temporally successive parts:
1) First, the ranking of the different images according to the feature under investigation. To help the test participants to perform the ranking, different indications were provided. They were told to sort the images by performing pairwise comparisons: they were instructed to look at the images already on the table and to compare each of them with the picture they have in hand. If the picture they have in hand has a higher value in terms of the property than one of the images on the table, then it should be located on its right side, and if not it should be located on the left side. All the images were put next to each other as a continuous depth cue line from low depth cue to high depth cue. Another indication was provided to the test participant: to perform the ranking, they were also told that they could first pre-sort the images by grouping them into different stacks of images corresponding to a group of images they think to have similar properties.

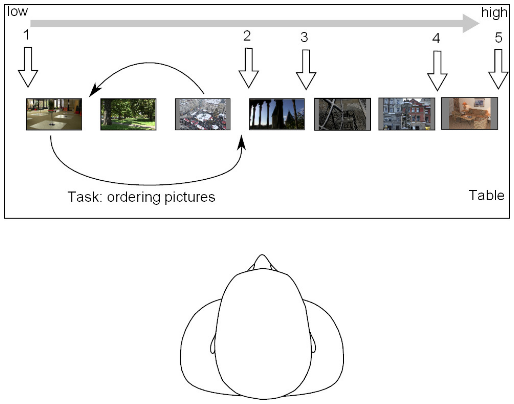


Figure 4.17: Ranking of printed images

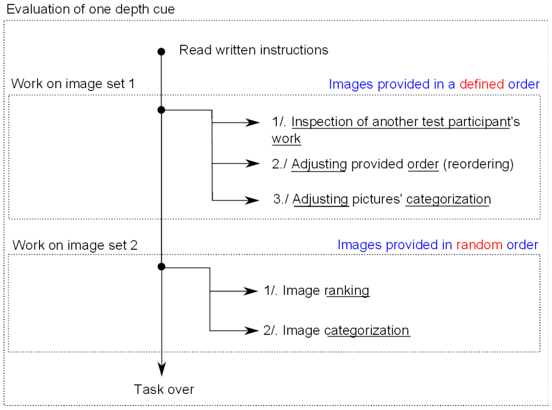

Evaluation of one depth cue
Read written instructions
Work on image set 1
Images provided in a defined order
1/. Inspection of another test participant's work
2./ Adjusting provided order (reordering)
3./ Adjusting pictures' categorization
Work on image set 2
Images provided in random order
1/. Image ranking
2/. Image categorization
Task over

Figure 4.18: Evaluation process sum-up

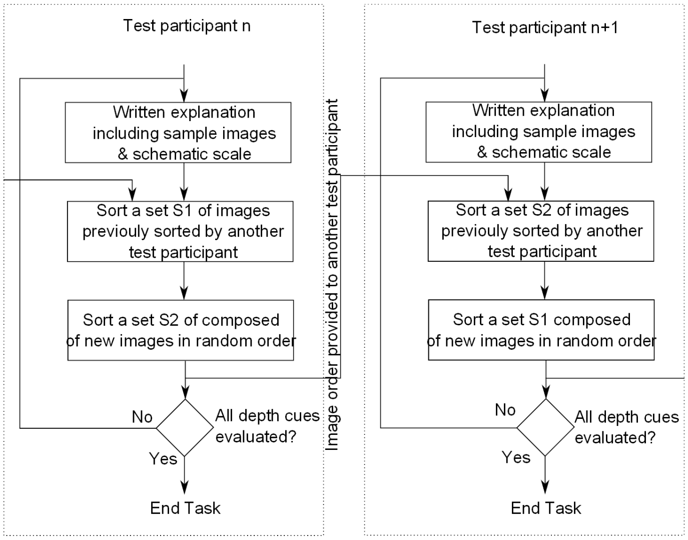

Test participant n
Written explanation including sample images & schematic scale
Sort a set S1 of images previouly sorted by another test participant
Sort a set S2 of composed of new images in random order
No
All depth cues evaluated?
Yes
End Task
Image order provided to another test participant
Test participant n+1
Written explanation including sample images & schematic scale
Sort a set S2 of images previouly sorted by another test participant
Sort a set S1 composed of new images in random order
No
All depth cues evaluated?
Yes
End Task

Figure 4.19: Sequential evaluation of the two image sets

2) Secondly, they had to add markers to group the pictures into four categories corresponding to the amount of perceived depth cues.

To train the participants on to the evaluation of the depth cues, the evaluation of each depth cue was divided into three parts as depicted in Figure 4.18. The first part consists in letting the test participant read the written instructions which includes a written description of the depth cue, five different pictograms which illustrate the depth cue, and six different examples (see subsection 4.6.2). After reading of the instructions, a set of 25 images was presented to the test participants; this set of images had a specific order and was the result of the ordering and marking of another test participant. The participant was informed about this, and was asked to (a) look at the provided order, which provides him/her additional examples of how to order the different images according to the considered depth cue, (b) adjust the order of the images according to what he/she thinks is the most appropriate order. It is only after these steps that the test participant was asked to order a new set of 25 images, which was provided in a random order. Once this task completed, they could start the evaluation of the next depth cue.

Figure 4.18 summarizes the different parts of the task. It corresponds to one iteration of the loop in Figure 4.18 and relates to the evaluation of one depth cue. The received picture order is as described previously and depicted in Figure 4.19, the result obtained from another test participant. As there two consecutive actions: re-ordering and ordering a set of image, and each set has 25 different images, some participants only saw one set of images during the reorder task and the other set in the ordering task. As the result of the ordering of one set of image was provided to the next participants, only half of the participants saw one set in the reordering task and the other half of the participants saw this same set in the ordering task. Therefore, later in the analysis two groups of participants will be considered depending on which set of image they had to reoder and order.

No time constraint was given to the test participant. On average, it took 17 min for each participant for ordering the 50 images for one scale.

The motivation behind the idea of ranking is to explicitly ask the test participants to compare the images to each other rather than choosing a quantitative score which can be a difficult task. A second advantage of the ranking approach is to always show the entire range of image's properties to the test participants. This can help them to define the order of two images, since they can see, at the same time, examples of extreme value of the considered property. The motivation behind grouping the images is to provide a way for test participants to report difficulties in performing the ranking between images, because they found them to have too similar properties, and then dispose of a way to report this by grouping images into categories. This categorization is different from traditional category rating since test participants have to carry out this task on a set of ordered images and only need to provide separations between groups of pictures. Moreover, they can see the entire range of the property when making this decision.

#### 4.6.1.2 Hypothesis and groups of test participants

Different questions on the methodology have been addressed and correspond to different groups of test participants. In total, 23 observers participated in the test.

To study the overall performance in terms of confidence intervals as compared to absolute category rating (ACR), all 23 test participants can be used. However, results can only be based on the data of the ranking and categorization of the images provided in a random order.

To study whether the ranking for the reordering task, help test participants to understand the scale under evaluation, the test participants were split into two groups: one group which received, as a first set of image to reorder, a set well defined by the test organizer. The second group received a first set of images to be reordered which was previously ordered by another test participant (as depicted in Figure 4.18). The first and second groups are composed of 16 and 7 test participants respectively. In the following, these two groups will be referred to as group 1 and group 2. The differences between the two groups will be studied.

To study whether reordering pictures results in more consistent ranking than ordering pictures, only test participants from group 2 can be used. This is done for avoiding the artificial higher consistency between test participants due to the fact that all participants from group 1 had to reorder the same image order.

Independently of who provided the image order to be reordered, 18 test participants had to order the first set of images, 5 had to reorder it. 19 test participants had to reorder the second set of images, 4 had to order it.

### *4.6.2 Description of the studied scale*

The description of the monocular depth cue scale was provided exactly according to Section 4.3 on the scale "the relative size". This depth cue has been chosen because of the difficulty to assess it and it thus challenges the proposed method. Moreover, this depth cue was addressed in the previous experiment and enabled to compare the proposed method to the traditional absolute category rating (ACR) methodology.

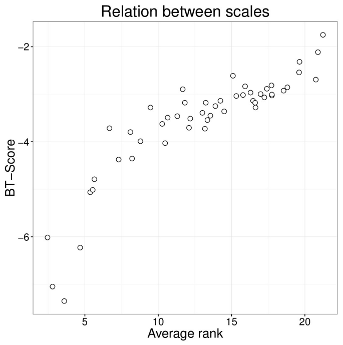


Figure 4.20: Relation between possible interpretation of the results.

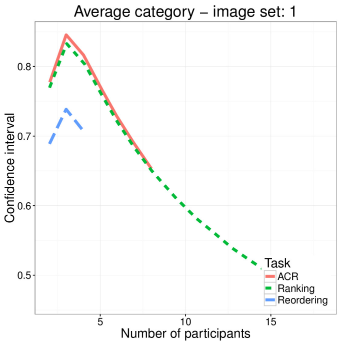


Figure 4.21: Confidence interval and test participant number

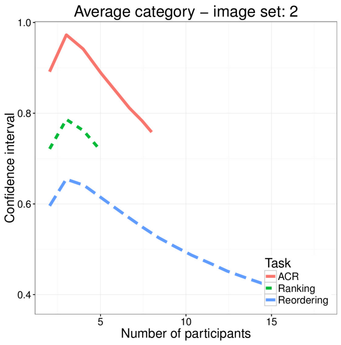


Figure 4.22: Confidence interval and test participant number.

## *4.6.3 Statistical analysis*

### 4.6.3.1 Interpretation of the data

Quantitative scores

Considering the different tasks within the test, alternative analysis can be performed, namely of:
1) Average rank provided by the test participants
2) Average of the category number to which the picture belongs.
3) The order provided by the test participants can be used to set up a pairwise comparison matrix describing for each position (i,j), the number of test participants who ranked an image i higher than an image j. Then, a model such as Bradley-Terry can be applied to obtain continuous scores [110].
This last type of analysis 3) is highly constrained since the ranking ensures transitivity. Then, if three images A, B, and C are ranked $A < B < C$, it implies that $A < C$. However, in a pair comparison experiment $A < B$ and $B < C$ does not necessarily imply $A < C$. Therefore, in the construction of the pairwise matrix, different options can be considered and it can be chosen to only use the comparisons between neighboring images (*option 3a*). Then, if $A < B < C$, only the relation $A < B$ and $B < C$ is used to set up the matrix since test participants were explicitly asked to compare neighboring images in the instruction of the test. Alternatively, it can be considered that test participants are only sure of how they ordered distant images (*option 3b*). In a ranking $A < B < C$, participants may not be sure if $A < B$ or $B < C$ but are at least sure of $A < C$. The number of images in between these two images A and C is a parameter which needs to be defined.

Comparison of the approaches

To compare the alternative approaches of interpreting the data, the Spearman correlation between the different scales is computed. It shows a correlation of 0.91 (see Figure 4.20), 0.99 and 0.89 respectively between options 1 and 3a, options 1 and 2, and options 2 and 3a. The high correlation between the different scales was expected since all the three proposed methods to analyze the data are highly related to each other, and there is a clear relation between how the test participants ranked the images, and how pairs of images are ordered compared to each other, and the category to which the picture belongs.
The approach for setting up the pairwise comparison matrix considering only pairs of images having a distance higher than a specific threshold was also applied (option 3b). When a threshold of 10 images between two images is defined, the analysis shows a Spearman correlation of 0.93 with options 1 and 2. When a threshold of 5 images is chosen, the correlation is as high as 0.97 with options 1 and 2.
The different ways to interpret the data thus appear to be similar in terms of the resulting scores.

#### 4.6.3.2 Inter-methodology analysis: ACR vs. Ranking

One hypothesis of the proposed approach is that ranking is an easier task for the test participants than rating the images on an absolute category scale. To check this hypothesis, it is possible to compare the confidence intervals between the category ratings, of the second ranking where the test participants had to order images which were provided to them in a random order, to the ACR scores of the study previously described in Section 4.5.3. The respective values are depicted in Figure 4.21. The confidence interval values for N test participants are computed based on average values of the confidence interval of all possible selections of N test participants between all the 23 test participants. This was done to avoid the confidence interval values of being too dependent of a particular selection of N test participants. As explained in section 4.6.1.1, two sets of images were studied. On the first set of images, the ordering of a new set of pictures which was not previously ordered by a test participant shows no improvement in terms of size of the confidence interval compared to the ACR method. On the second set, an improvement of the size of the confidence interval of 0.18 can be seen and is depicted in Figure 4.22. The confidence interval using the ACR-methodology of the second set of images appears to be larger than for the first set of images. It appears that this set was more difficult to evaluate, and the ranking methodology may have simplified the task to the test participants.

#### 4.6.3.3 Intra-methodology analysis

A first hypothesis H1 is that reordering the first set of image provided more stable results since the test participants were able to check and correct mistakes made by the other test participants. A second hypothesis H2 is that a solution from another test participant can help the test participants to better understand the scale under evaluation. To test the latter hypothesis, the test participants were divided into two groups (see Subsection 4.6.1.2): The test participants belonging to the first group had to reorder the second set of images. This explains the much smaller confidence interval of the reordering task in Figure 4.22: all participants received the same order and image categorization.

Effect of choosing the set to reorder on user consistency

Here, the aim was to study whether providing a well-controlled image order to be reordered has a positive effect on how test participants understand the target scale; the inter-correlation between the test participants from group 1 and 2 is compared regarding their ability to provide consistent results in ordering a new set of images. For each observer of group 1 or 2, the Spearman correlation between his/her ordering and the ordering of the other test participants of the same group was computed. The distribution of the inter-correlation values between test participants of group 1 and 2 is then compared using a Kruskal-Wallis one-way analysis of variance, showing a p-value of 0.068. Therefore, no statistical differences (at 95% confidence) can be found between the agreements of the test participants during the second task, regardless of whether they received a set of image to be reordered defined by us or by another test participant during the first task. To quantify the differences of agreement between the two groups of test participants, the Fleiss' kappa test [129] was computed and was found equal to 0.31 for the group 1, and 0.11 for the group 2. Both values are low, but show an improvement of the agreement for group 1, which is found to be "fair" whereas it is "slight" for group 2. Even if these agreements are low, in both cases the test shows that the agreement between the observers is not accidental at 95% confidence: ($p < 0.01$, z=27) for the first group and (p=0.01, z=2.5) for the second group.

Unfortunately, this does not enable a strong conclusion on the hypothesis H2 to be drawn. Moreover, it should be mentioned that the classification between the image's order provided by another test participant or the author is not enough to differentiate the "quality" of the image rank to reorder.

These results show, however, the expected tendency.

User consistency between reordering and ordering tasks

To study if the process of reordering images enables to have a higher inter-test participant agreement between reordered images and ordered images, statistical tests will be done. The agreement between test participants of group 2, as defined in the previous subsection, is compared to the agreement of these same test participants when doing the ordering of a set of images provided in a random order. Similarly to the analysis in the previous subsection, the distribution of the Spearman inter-correlation between the rankings of each test participant compared to the other test participants when doing the *reordering* task is compared to the distribution of the Spearman correlation values between the test participants when doing the *ordering* task. Due to the permutation between image sets received by each test participant: one test participant reorders a set of images and orders the second set of images, the second set being provided to the next test participant as a set to reorder (see Figure 4.18). Three test participants were asked to reorder the first set of images and order the second set, and four test participants were asked to do the opposite. To study the differences of consistency between test participants during the ordering and reordering tasks, a Kruskal-Wallis one-way analysis of variance is used to compare the distribution of the observer inter-correlation depending on the task to perform: ordering or reordering. It shows a p-value of 0.44 and 0.13 respectively for the task of reordering image set 1 and ordering image set 2, and reordering image set 2 and ordering image set 1. In both cases, no statistical differences at 95% confidence can be found between the agreement of the test participants, depending on whether they had to order or reorder the images.

To further analyze the differences between tasks (ordering or reordering) in terms of user consistency, the Fleiss'es kappa test was computed and result values can be found in Table 4.4. In every case the agreement values between test participants are low and are all classified as "slight" agreement. A small increase of agreement between the test participants can however be seen. This slight increase of consistency is also visible in Figure 4.21 in terms of confidence interval size.

Based on these results, the hypothesis H0 has to be rejected since no statistical differences in the agreements between test participants after ordering or reordering can be observed. However, only a small number of test participants could be included in the statistical analysis of this hypothesis. A small difference in the expected direction of increasing the consistency between test participants is visible. Increasing the number of test participant may contribute to better reveal the increase of agreement.

| | Reorder set 1 | Reorder set 2 |
|---|---|---|
| ordering task | 0.091 | 0.091 |
| reordering task | 0.10 | 0.12 |

Table 4.4: Fleiss'es kappa depending on task and image group

### *4.6.4 Limits*

The proposed method was designed to make the task of evaluating the relative size depth cue easier for the test participants. A clear limitation of this method is that it can mainly be applied to evaluate characteristics in images. It is believed that using printed images enabled the test participants to freely examine all the images and quickly change from one image to another always having a view on the overall set of images. Such approach is rather unpractical to implement with videos.

A second limitation of the test is that the current method does not provide any time constrains to the test participant. This does not allow the overall length of the test to be controlled. The time taken by each test participant was rather constant, but this issue needs to be considered in future tests.

### *4.6.5 Discussion*

On a side note, it can be observed in Figures 4.21 and Figure 4.22, that the image set 1 was used by most participants (18/23) during the ordering task while the image set 2 was used by most participants during the reordering task (18/23). This is not balanced: there is not an equal number of participants who saw the first set of images during the ordering and reordering task. However, even with the limited number of participants who used the first set of images as a reordering task, it is possible to observe as analyzed in the previous sections, an increase of consistency between participant raking. Therefore whether the first set or second set of images was used during the reordering task does not appear to have a too strong effect on the goal of increasing inter-participant agreement on images ordering.

### *4.6.6 Analysis per depth cue*

As described in the previous section 4.6, in the process of the test, test participants were requested for each depth cue to reorder a set of images before being asked to order a new set of images. Figure 4.23 depicts different cases of reordering between the provided rank and the rank returned by the test participant. It can be seen that some test participant only slightly altered the provided order and other test participants made big changes. On Figure 4.23.d, the particular case of one test participant who completely altered the order of the pictures is depicted. On Figure 4.23.e, the order provided by this test participant is reordered by another test participant which also strongly disagree with the provided order and provides a new raking which is in a better agreement with the other test participants (Figure 4.23.f ). Such analysis may be used to identify observers who rates differently. To determine how two observers agree, it is proposed to determine the number of times images are ordered differently from each other compared to a specific threshold (see Figure 4.23.f). Figure 4.24 depicts the average over the test participant of the differences between the rank they received and the rank they returned after reordering with different values of threshold. An ANOVA shows that test participant made significantly more changes for the linear perspective and relative size than for the interposition ($F = 41.96, p < 0.01$). There are two potential explanations for the difference of agreement between the test participants for each depth cue: 1) the test participant has well understood the scale and converged to the final solution resulting in a high agreement between participants. 2) The opposite interpretation can also be possible; the test participants may have found the scale too difficult to evaluate and chose not to change the provided image rank since they did not know how to order the images. To study, which is the most likely hypothesis, the agreement between the test participant was measured on the second task of the experiment. In this second task, the test participants were asked to order images from a random order. If the test participants agree on the scale, there should be only small differences between the order provided by each person compared to each other. Figure 4.25 depicts the average difference between observers per depth cue scale on the two sets of images. The evaluation of the two image sets differs on the first part of the evaluation: whether test participants had to reorder a set of images defined by the author (image set 1), or by another test participant (image set 2). In case of the second image set, no differences can be observed between the scales. However, in the first image set, both scale “Interposition” and “Relative Size” appear to provide a higher agreement between the test participant than “Linear perspective”. Kruskal-Wallis one-way analysis of variance shows that the agreement for Interposition and Relative size are significantly different from the agreement on Linear perspective (p-values are respectively equal to 0.0378 and 0.0113). But the differences of agreements between Interposition and Relative Size are not statically different (p=0.238). These results show that after reordering a set of images provided by the author, test participants were more consistent in the Interposition scale than the Linear Perspective scale. This result goes into the direction that test participants better understood the Interposition than the Linear Perspective scale, and explains the higher agreement between participants in the first reordering task. This is visible in Figure 4.24. However, this may not be the only factor involved, since 1) the Relative Size shows a similar agreement between the received order and the result of the reordering converged to the same comparison on the Linear Perspective. 2) In the ordering task, the agreement between test participants is similar between Interposition and Relative Size. One possible hypothesis could be a variation of difficulties between image set as a new factor to take into account in the analysis.

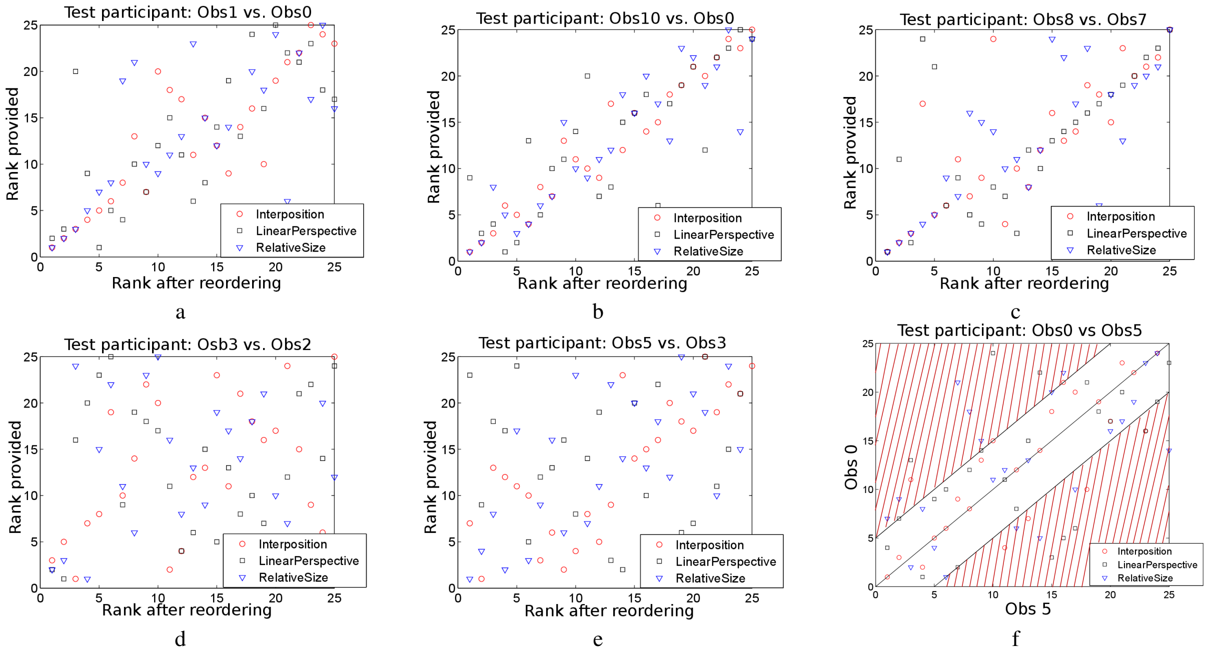


Figure 4.23: Example of relation between the provided ranking and rank returned by a test participant.

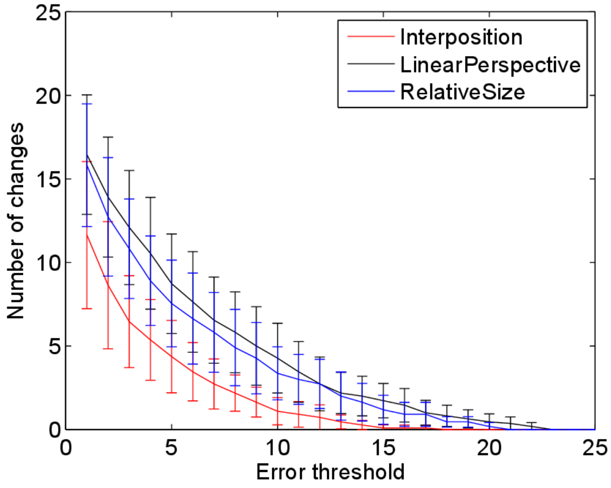


Figure 4.24: Average differences between ranking received order and order provided by the participants after reordering considering different error thresholds. The average is performed across all participants.

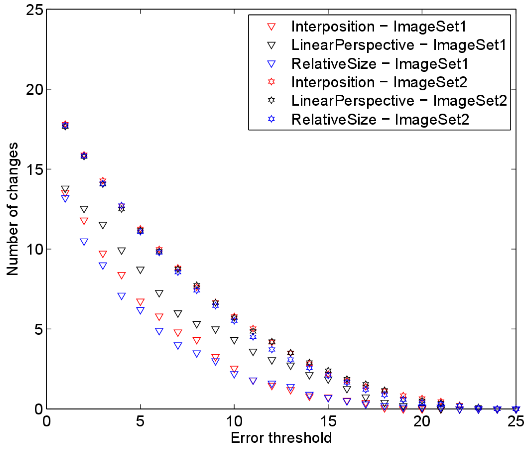


Figure 4.25: Average differences between ranking of different test participant in the ordering from random order task on the two sets of images.

### *4.6.7 Conclusion*

In this section, a new method based on rankings was presented to evaluate complex dimensions such as the relative size depth cue in natural images. Results have shown that the proposed method provided either as consistent results as the ACR methodology, or even more stable results in case of a difficult image set. The proposed methodology is divided into two parts: first reordering a set of images from another test participant and then ordering a new set of images. Such an approach has shown to help the test participants to better understand the scale, and an improvement of the consistency of the result has been observed. This improvement was, however, only close to being significant. The difference of consistency between reordering and ranking was also considered, a small improvement was observed but was not significant and will require more extensive tests to be proven.

While all test design hypotheses could not be proven to be statistically effective, the approach based on rankings appeared to improve stability of results when a difficult feature has to be evaluated. It is proposed to be applied to the study of monocular depth cues in natural images.

**Listing 4.7: Conclusion on the research questions**

```
1. Showing the result of another test participant providing an example of results was
    found to improve the agreement between test participants but was only close to reach
     significance at a 95% confidence.
2. Using ranking, it was possible to decrease the size of the confidence interval
    compared to ACR when the same number of participants are considered.
3. An improvement of the agreement between test participants was observed, when each
    test participants are asked to adjust the work of another test participant.  However
    , the improvement was not significant.
```

## 4.7 Key contributions

In this chapter it was addressed how depth can be evaluated. Moreover it was considered how natural images can be characterized on different scales described by the different monocular depth cues. The main contributions are:

- First it was shown that the evaluation of perceived depth is difficult, and there is a high variance between test participants.
- The second and main contribution of the chapter is the proposal of a method to perform the evaluation of monocular depth cues on natural images. Most of the state of the art methods have focused on specifically designed signal, but have not considered natural images. The use of these kinds of images has brought new challenges: first, the need to characterize them along different axis by providing definition and instruction on how to quantify the different monocular depth cues scales. Secondly, it requires to develop new subjective methods for the evaluation of these depth cues. The newly proposed method was compared with the traditional absolute category rating approach, and have enabled to reach higher consistency between test participants.
- Finally, the image datasets and scores were distributed as Open Source database including newly created images and evaluated across different axis corresponding to different depth cues.

# Chapter 5
# Algorithms for depth evaluation

## 5.1 Introduction

The previous chapter has addressed the question of evaluating monocular and binocular depth cues in natural images. In this chapter it is proposed to present several algorithm which were designed for the prediction of 3D images properties considering both monocular and binocular depth cues. First binocular depth cues will be addressed, then several monocular indicator will be described. Finally, considering that these algorithm are error prone, the question of indicator reliability will be addressed and different methods will be presented on how to determine cases of failure and take this information into account in the process of depth cue pooling.

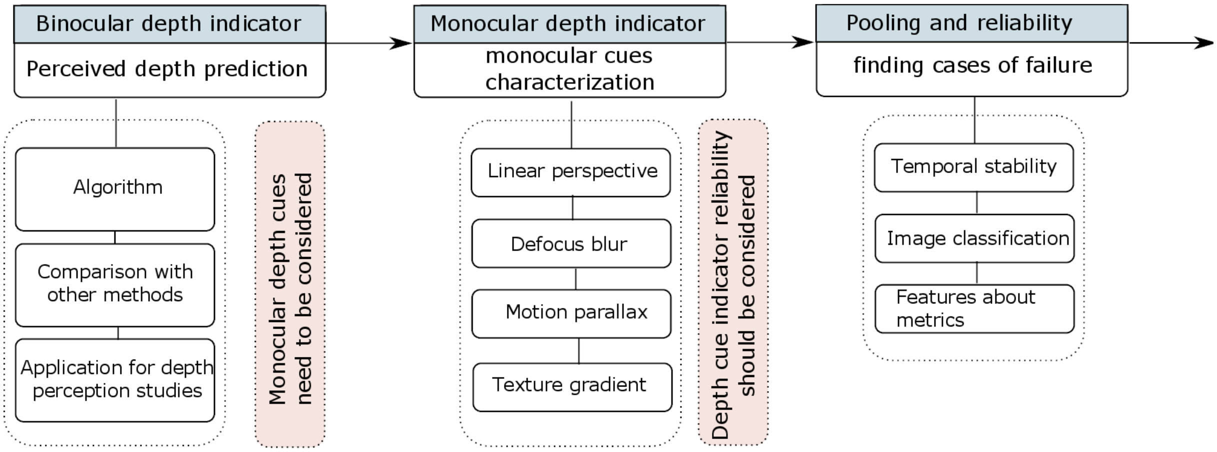

Binocular depth indicator
Perceived depth prediction
Algorithm
Comparison with other methods
Application for depth perception studies
Monocular depth cues need to be considered
Monocular depth indicator
monocular cues characterization
Linear perspective
Defocus blur
Motion parallax
Texture gradient
Depth cue indicator reliability should be considered
Pooling and reliability
finding cases of failure
Temporal stability
Image classification
Features about metrics

Figure 5.1: Different studied items

## 5.2 Instrumental characterization

The characterization of video characteristics has been addressed in the context of 2D video sequences. In this case, the spatial and temporal complexity indicators SI and TI were defined in the ITU-T Recommendation P.910 [116]. These indicators aim to provide information on the complexity of the image taken individually, and the temporal variation of the video content. These indicators are defined as folllows: Let $Y_n(i,j)$ be the value of the luminance of the pixel at the position (i, j) in the frame n. A sobel filter is applied to the frame to determine the gradient along the vertical and horizontal direction (Eq. 5.1 and Eq. 5.2).

$$\begin{aligned} Gv_n(i,j) = & \, 1\cdot Y_n(i-1,j-1) - 2\cdot Y_n(i-1,j) - 1\cdot Y_n(i-1,j+1) \\ & + 0\cdot Y_n(i,j-1) + 0\cdot Y_n(i,j) + 0\cdot Y_n(i,j+1) \\ & + 1\cdot Y_n(i+1,j-1) + 2\cdot Y_n(i+1,j) + 1\cdot Y_n(i+1,j+1) \end{aligned} \tag{5.1}$$

$$\begin{aligned} Gh_n(i,j) = & \, 1\cdot Y_n(i-1,j-1) + 0\cdot Y_n(i-1,j) + 1\cdot Y_n(i-1,j+1) \\ & - 2\cdot Y_n(i,j-1) + 0\cdot Y_n(i,j) + 2\cdot Y_n(i,j+1) \\ & - 1\cdot Y_n(i+1,j-1) + 0\cdot Y_n(i+1,j) + 1\cdot Y_n(i+1,j+1) \end{aligned} \tag{5.2}$$

Then, the norm of the gradient is computed (Eq. 5.3)

$$G_n(i,j) = \sqrt{Gv_n(i,j)^2 + Gh_n(i,j)^2} \tag{5.3}$$

The value of SI is defined as the maximum value of the standard deviation over the time of the Sobel-filtered frames (Eq. 5.4)

$$SI = max_{time}(StdDev_{space}(G_n(i,j))) \tag{5.4}$$

The value of TI is defined as the maximum value of the standard deviation of the difference of pixel luminance values between two consecutive frames in a specific window (Eq. 5.5).
Let defined $TI_n(i,j) = Y_n(i,j) - Y_{n-1}(i,j)$ with $Y_n(i,j)$ the value of the luminance of the pixel at the position (i,j) in the frame n.

$$TI = max_{time}(StdDev_{space}(TI_n(i,j))) \tag{5.5}$$

The information about content properties can then be used to perform content selection, as described in ITU-T Recommendation P.910. Figure 5.2 illustrates the selection process of 2D video sequences for a subjective test. Video sequences must have different properties in order to avoid result for a too narrow set of video. Such characterization, without removing the need of subjective inspection, enables to identify sources with different spatial and temporal complexities from a large database.

However, these measurements have their limits: considering these definitions, one can observe that if a video sequence is highly textured then the value of SI will be high. But one can also see that the high amount of texture will have a strong impact on the value of TI: if there is a high amount of texture, even little motion will create high variations of differences of luminance of consecutive frames, and thus high values of TI. This is illustrated in Figure 5.2, where the distribution between the SI and TI values appears correlated. Nevertheless, as long as their limits are known, such algorithm can still be useful in case of image selection for large database, or in case of image and video quality prediction to have a characterization of the 2D videos.

In the context of 3D video sequences, such indicators need to be defined for describing the “3D effect”. The depth effect depends on different factors which are related to both the binocular and monocular vision. This chapter will detail the work conducted on the characterization of 3D video properties though different depth cues: the monocular and the binocular ones.

## 5.3 Background

Before describing the contributions made in the domain of content characterization and perceived depth estimation, it is useful to describe the different kinds of algorithm which are needed by the proposed algorithms. Different types of features were required such as depth maps from different kinds of depth cues, segmentation of images, vanishing line

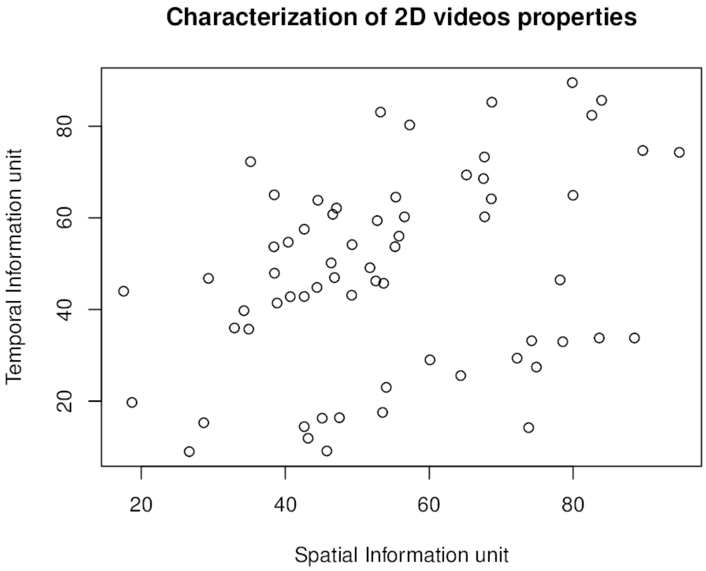


Figure 5.2: Selection of source sequence based on spatial and temporal complexity. Each points corresponding to a different video having different spatial and temporal complexity. The video depicted in this figure are the ones listed in Table 4.2 in Section 4.4 and will be evaluated regarding their 3D properties in this chapter.

extraction, etc.

One of the key features required to estimate depth perception in stereoscopic images are the binocular depth maps. These depth maps could be either obtained during the capture process using depth cameras, or estimated in a post process step. The second case, depth estimation, is a particularly challenging task but frequently required since currently most video recording is performed without depth cameras. In this section methods for estimating dense depth map from stereoscopic images will be described.

### *5.3.1 Depth maps from stereoscopic videos*

The estimation of depth from stereoscopic images is no easy task. It consists of estimating the correspondences between corresponding pixels of two stereoscopic images, and is similar as the problem of estimating dense optical flow (motion). In this subsection, different methods will be presented to solve the problem of stereo correspondence.

#### 5.3.1.1 The Horn & Schunck algorithm

Amongst the current best performing approaches, many of them are based on the work conducted by Horn & Schunck in 1981 [130] who replaced the traditional block-based and feature-based matching algorithms by a minimization problem of a general energy function:

$$argmin_{u,v} \quad E = \int |\nabla u|^2 + |\nabla v|^2 dxdy + \lambda \int \rho(u,v)^2 \tag{5.6}$$

These types of approaches became popular, since they are highly parallelizable and then suitable for GPU programming. The equation is divided in two parts. The second part of the equation is called “data-term” and characterizes how close the solution (u, v) for horizontal and vertical motion is close to the ideal solution. However, considering that the problem is ill-posed, it is not possible to find a solution considering only the data-term. Assumptions about the motion must be made; therefore a second part is added to the equation and is called the “regularization term”. This is the first part of the energy function and define that the motion variations should be smooth and therefore the motion gradient must be as small as possible. A coefficient is introduced between the regularization and data term to weigh the relative importance of both parts of the equation. Estimating motion is then reduced to finding the motion vectors (u,v) such as the difference between the two images is as small as possible and having as smooth as possible variation of the motion vectors.

The next step is to find a solution. The luminance corresponding to the position $(x,y)$ in the frame t is noted $I(x,y,t)$. The corresponding pixel in the second image is $I(x+u,y+v,t+1)$. Using Taylors expansion around 0 and assuming that the motion is small:

$$I(x+u,y+v,t+1) \approx I(x,y,t) + \frac{\partial I}{\partial x}(x,y,t) \times u + \frac{\partial I}{\partial y}(x,y,t) \times v + \frac{\partial I}{\partial t}(x,y,t) \times dt \tag{5.7}$$

With dt = 1. The data-term can then be rewritten:

$$\begin{aligned} \int \rho(u,v)^2 dxdy &= \int (I(x+u,y+v,t+1) - I(x,y,t))^2 dxdy \\ &= \int (\frac{\partial I}{\partial x}(x,y,t) \times u + \frac{\partial I}{\partial y}(x,y,t) \times v + \frac{\partial I}{\partial t}(x,y,t))^2 dxdy \end{aligned} \tag{5.8}$$

Using the calculus of variation, (u, v) is found that it must verify the following Euler-Lagrange equations:

$$\frac{\partial I}{\partial x}(x,y,t) \times (\frac{\partial I}{\partial x}(x,y,t) \times u + \frac{\partial I}{\partial y}(x,y,t) \times v + \frac{\partial I}{\partial t}(x,y,t)) - \lambda \times (\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) = 0 \tag{5.9}$$

$$\frac{\partial I}{\partial y}(x,y,t) \times (\frac{\partial I}{\partial x}(x,y,t) \times u + \frac{\partial I}{\partial y}(x,y,t) \times v + \frac{\partial I}{\partial t}(x,y,t)) - \lambda \times (\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}) = 0 \tag{5.10}$$

However in order to be able to apply the Taylor expansion, it was assumed that the motion is small and close to 0. This is not realistic. It is then suggested that if an estimation of $(u,v)$ noted $(\hat{u},\hat{v})$, is known it would be possible to determine small adjustment of the motion vector, noted $(du,dv)$.

$$\hat{I}(x,y,t+1) = I(x+\hat{u},y+\hat{v},t+1) \tag{5.11}$$

$$\hat{I}(x,y,t+1) \approx \frac{\partial \hat{I}}{\partial x}(x,y,t+1) + \frac{\partial \hat{I}}{\partial y}(x,y,t+1) + \hat{I}(x,y,t+1) \tag{5.12}$$

Then the data term is equal to:

$$\int \rho(u,v)^2 dxdy = \int (\frac{\partial \hat{I}}{\partial x}(x,y,t+1) + \frac{\partial \hat{I}}{\partial y}(x,y,t+1) + \hat{I}(x,y,t+1-I(x,y,t)))^2 dxdy \tag{5.13}$$

Let $\hat{I}_x = \frac{\partial \hat{I}}{\partial x}$, the previous Euler-Lagrange equation becomes:

$$\begin{aligned} (\hat{I}_x du + \hat{I}_y dv + \hat{I} - I) \cdot \hat{I}_x - \lambda(du_{xx} + du_{yy}) = 0 \\ (\hat{I}_x du + \hat{I}_y dv + \hat{I} - I) \cdot \hat{I}_y - \lambda(dv_{xx} + dv_{yy}) = 0 \end{aligned} \tag{5.14}$$

Using finite difference method, the Laplace operator can be express as:

$$(\Delta u)_{i,j} \approx (u_{i-1,j} - 2 \cdot u_{i,j} + u_{i+1,j}) + (u_{i,j-1} - 2 \cdot u_{i,j} + u_{i,j+1}) \tag{5.15}$$

The Euler-Lagrange equation becomes:

$$(\hat{I}_x du + \hat{I}_y dv + \hat{I} - I) \cdot \hat{I}_x - \lambda (\Delta du)_{i,j} = 0$$
$$(\hat{I}_x du + \hat{I}_y dv + \hat{I} - I) \cdot \hat{I}_y - \lambda (\Delta dv)_{i,j} = 0 \quad (5.16)$$

It is then possible to use the Jacobi method to define a sequence which will converge to an estimate of the adjustment of the motion vector $(du, dv)$.

$$du^{n+1}_{i,j} = du^n_{i,j} - \frac{\hat{I}_{x,i,j}(\hat{I}_{x,i,j} du^n_{i,j} + \hat{I}_{y,i,j} dv^n_{i,j} + \hat{I}_{i,j} - I_{i,j})}{\lambda + I^2_{x,i,j} + I^2_{y,i,j}}$$
$$dv^{n+1}_{i,j} = dv^n_{i,j} - \frac{\hat{I}_{y,i,j}(\hat{I}_{x,i,j} du^n_{i,j} + \hat{I}_{y,i,j} dv^n_{i,j} + \hat{I}_{i,j} - I_{i,j})}{\lambda + I^2_{x,i,j} + I^2_{y,i,j}} \quad (5.17)$$

With the initial value of the adjustment vector: $(du^0_{i,j}, dv^0_{i,j})$ null.
However, the method still requires an estimation of the value of the motion vector $(\hat{u}, \hat{v})$. The proposed algorithm can only determine the adjustment of the estimation of the motion vector $(\hat{u}, \hat{v})$. This initial motion vector $(\hat{u}, \hat{v})$ cannot be null since the hypothesis requires that the adjustment vector is close to 0 and the motion vectors themselves will not be null.
To solve this issue, as depicted in Figure 5.3, the motion estimation is done at multiple resolutions. At a very low resolution, the motion is small. It is then possible to assume that an initial estimation of the motion vector at this very low resolution is null: $(u, v) \approx (0, 0)$. Then it becomes possible to estimate a refinement of the motion vector: (du, dv). The refinement is then used as initial value of the motion at a higher resolution. A refinement is again computed and will be used again as an initial value of motion for higher resolution motion estimation. This is done until the motion estimation is made on the finest resolution.

#### 5.3.1.2 Total variation - $\ell_1$

Improvements have been proposed to the Horn & Schunck algorithm. One of the issues comes from the regularization term which ensure local smoothness of the solution. However, the local smoothness may not always apply, and abrupt variation of motion within spatial location in the image may happen. To tackle this problem, the $\ell_1$-norm can be used as an alternative to the $\ell_2$-norms proposed by Horn & Schunck, which enables larger difference:

$$argmin_{u,v} \quad E = \int |\nabla u| + |\nabla v| dxdy + \lambda \int |\rho(u, v)| dxdy \quad (5.18)$$

To solve this equation, an efficient approach is the use of Primal-Dual optimization algorithms. The problem is divided into the following two subproblems:

- One minimization problem: decrease of the data term.
- One maximization problem: increase the smoothness of the solution.

Let $p$, be the dual variable of the motion vector $(u, v)$. The convex conjugate of the total variation term, $|\nabla(u, v)|$, is the function to optimize:

$$F^*(p) = sup_{(u,v)} \{\langle \nabla(u, v), p \rangle - |\nabla(u, v)|\} \quad (5.19)$$

As demonstrated by Werlberger [131], the solution to the equation 5.18 can be found after convergence of the following numerical series:

$$p^{n+1} = prox_P(p^n + \sigma \nabla(\bar{u}^n, \bar{v}^n)$$
$$(u^{n+1}, v^{n+1}) = shrink((u^n, v^n) - \tau div p^{n+1})$$
$$(\bar{u}^{n+1}, \bar{v}^{n+1}) = 2(u^{n+1}, v^{n+1}) - (u^n, v^n) \quad (5.20)$$

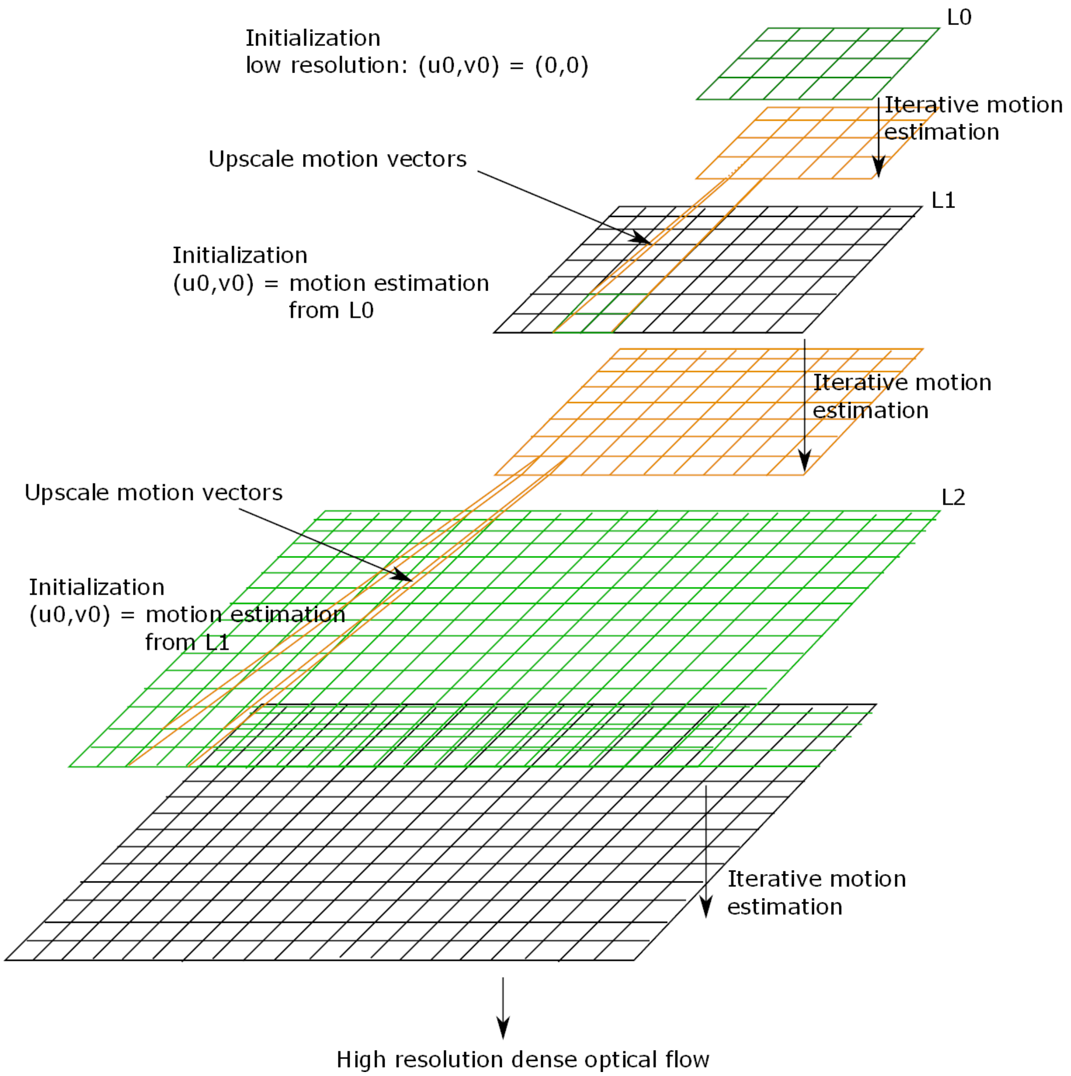


Figure 5.3: Pyramidal approach for dense motion estimation.

With the operator *prox* defined as the projection of the vector to a unit ball, and *shrink*s is a conditional function defined as in Table 5.1.

| Condition | Threshold check | Updating |
|---|---|---|
| $\rho(u) > 0$ | $\rho(\hat{u}) < -\tau\lambda\nabla I$ | $u = \hat{u} + \tau\lambda\nabla I$ |
| $\rho(u) < 0$ | $\rho(\hat{u}) > \tau\lambda\nabla I$ | $u = \hat{u} - \tau\lambda\nabla I$ |
| $\rho(u) = 0$ | $\|\rho(\hat{u})\| <= \tau\lambda\nabla I$ | $u = \hat{u} - \nabla I \frac{\rho(\hat{u})}{\|\nabla I\|^2}$ |

Table 5.1: Definition of the *prox* operator

Similarly as before, to initialize the series, the value of $(u^0, v^0)$ needs to be known. To get, an estimate of it, a multiple resolution approach is again performed.

#### 5.3.1.3 Further refinements

To enable further discontinuities around edges, the regularization term can also be defined as depending on a function of the motion gradient. This can enable decreasing the influence of the regularization term on the borders [132].

$$R = \int \varphi(|\nabla(u,v)|)dxdy \tag{5.21}$$

In case of anisotropic flow regulation, the cost function of the regularization term will also consider the orientation of the edges to decrease the influence of the regularization on the direction crossing the edge but will increase the influence of the regularization in the direction parallel to the edges [133].

$$R = \int tr(\varphi(\nabla(u,v)\nabla(u,v)^T))dxdy \tag{5.22}$$

In the particular case of depth estimation it is also possible to have isotropic image driven regularization: the cost function of the regularization term is weighted by the image gradient [134] and information obtained from monocular depth cues, such as blur [135].

$$R = \int G(I) \times \nabla(u,v)dxdy \tag{5.23}$$

In the context of this thesis, an anisotropic Huber-$\ell 1$ regularization was used to estimate the depth maps [136].

#### 5.3.1.4 Application to depth estimation & discussion

The application of algorithms designed for motion estimation to depth estimation can be problematic. In the case of stereo matching, larger movement between objects composing the scene can be observed compared to the kind of movement which happens in the temporal aspect in the case of motion estimation. This results in large discontinuities in the optical flow which contradict the smoothness constraint defined by the regularization term. Figure 5.4 depicts an example of the depth map produced using such approach. It is clearly visible that edges in the depth map do not show abrupt transition as they should have and appear over-smooth. This is due to the local continuity assumption as previously explained.

The topic of stereo-matching is a difficult one, and would have required a large amount of work beyond the scope of this thesis. Different methods have been considered to estimate the depth map. The University of Middlebury provides an extensive benchmark of stereo-matching algorithms [137]. However most of the algorithms described do not provide an implementation. Some implementation of stereo-matching algorithms can be found, but the ones which have been tested during the research work did not appear to handle all the diversity of content addressed in the thesis. Therefore, even if the Huber-$\ell 1$ dense optical flow used in this work perform lower on the Middlebury database, it was selected as a solution to estimate dense depth map since it was able to address the large variety of content used in the thesis. Its limits have been clearly identified: the sharpness of the edges of the depth map is far from optimal, and would not be good enough for a context of depth-based image rendering, but the performance can be sufficient for the context of this work on quality and depth characterization where distribution of disparity values are studied.

### *5.3.2 Depth estimation from monocular depth cues*

In addition to stereo-matching, many approaches have been designed to estimate the depth of images using monocular depth cues. Different methods to address this issue will be described in this section.

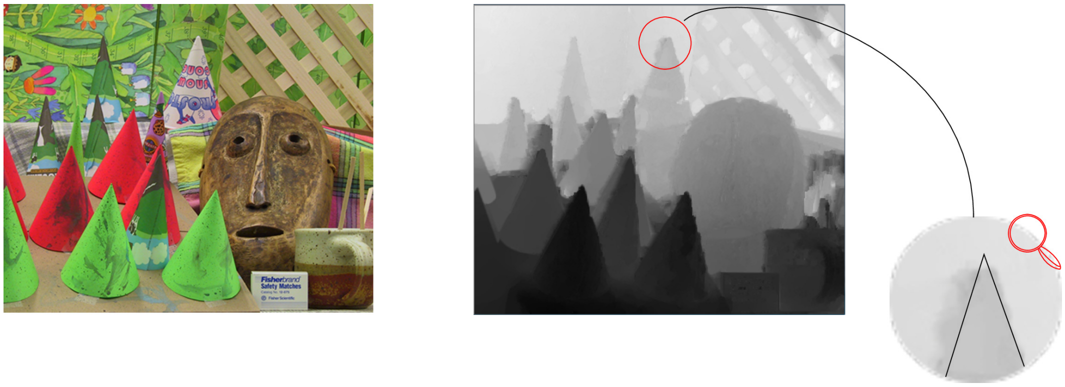


Figure 5.4: Estimation of depth map using a pair of images.

#### 5.3.2.1 Defocus blur

As detailed in chapter 2.2.2, defocus blur can be used as a measure of depth. If the focal length of the lens, and the distance between the point of focus and the lens is known, it is possible to evaluate absolute but unsigned values of distance between objects and the focus point based on the amount of blur. In this section, the issue of the evaluation of dense blur map and their conversion to depth will be addressed.

One of the issues about the estimation of blur is the ability to estimate the defocus blur on non-textured areas. Across edges, different methods have been proposed. Amongst them, methods have been proposed to evaluate blur by measuring the slope of the edge's gradient [138], the effect of the convolution by a Gaussian kernel on a picture [139], the distribution of the DCT coefficients [140], etc. However these only enable to have localized measures of blur on edges. On a non-textured area, such as depicted in Figure 5.5, these metrics will fail to differentiate between absence of a blurred image and image without edges. In most cases, this issue can be addressed due to the granularity of the output which is expected: a global indicator of blur and not blur measure at every location of the picture. In that case, the sharpness of the picture is usually sum-up to the maximum sharpness measured in the picture. However in the context of this study it is a per-pixel measure of blur which is expected in order to obtain a dense depth map from blur measurements.

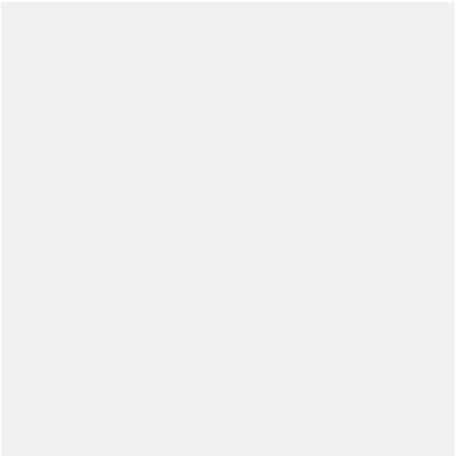


Figure 5.5: Blurred picture, or not?.

To obtain dense blur maps, the method proposed by Zhuo [141] was used. In order to get an estimate of the blur at every location within the picture, the overall process of blur from defocus depth cue evaluation is divided into two steps:

- Sparse blur estimation based on areas having edges.
- Interpolation of blur values between blur values measured on edges.

To get an estimate of the blur of edges, the algorithm apply a canny edge detector to determine the areas of the picture where blur estimation can be performed reliably. Figure 5.7 depicts an example image of a tree branch where only edges of the tree branch can be used to evaluate the amount of blur. The background of the picture is uniform and cannot be used to estimate reliably the defocus blur. To measure the amount of blur on edges, the blurred picture can be modeled via the equation 5.24, where $I_b$ is the blurred picture, $I_o$ is the non-blurred version of the picture, and $g_\sigma$ is the defocus Gaussian blur which needs to be estimated.

Figure 5.6: Orginninal image to process

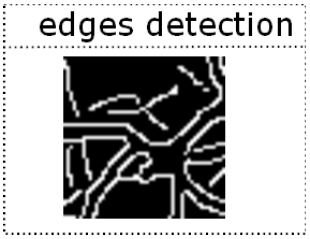


Figure 5.7: Edge detection

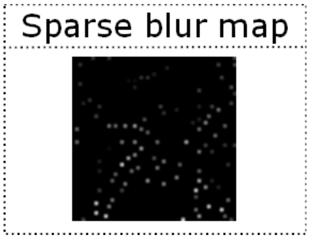


Figure 5.8: Sparse blur map

$$I_b = I_o * g_\sigma \tag{5.24}$$

To estimate the amount of blur, a Gaussian blur $g_{\sigma_0}$ is applied to image 5.25, and the effect of re-blurring the picture is evaluated by computing the gradient of edges identified previously. To simplify the notations, the following equations will only address a one-dimensional picture, but the results can be extended to further dimensions. Equation 5.28

$$I_{rb} = I_o * g_\sigma * g_{\sigma_0} \tag{5.25}$$

$$\begin{aligned} \nabla I_{rb} &= \nabla(I_o * g_\sigma * g_{\sigma_0}) && (5.26)\\ &= \nabla((A \cdot u(x) + B) * g_\sigma(x) * g_{\sigma_0}(x)) && (5.27)\\ &= \frac{A}{\sqrt{2\pi(\sigma^2+\sigma_0^2)}} exp\left(-\frac{x^2}{2(\sigma^2+\sigma_0^2)}\right) && (5.28) \end{aligned}$$

The ratio of the gradient norm before and after filtering by the Gaussian blur $g_{\sigma_0}$ provides the relation described by eq. 5.29. The ratio will be maximized at the edge locations, $x = 0$, which simplifies the equation to eq. 5.30. Therefore, on the edges it is possible to determine the properties of the blur as described in eq. 5.31. This provides the sparse blur map depicted in Figure 5.8.

$$\frac{|\nabla I_o(x)|}{|\nabla I_{rb}(x)|} = \sqrt{\frac{\sigma^2+\sigma_0^2}{\sigma^2}} exp\left(-\left(\frac{x^2}{2\sigma^2} - \frac{x^2}{2(\sigma^2+\sigma_0^2)}\right)\right) \tag{5.29}$$

$$R = \frac{|\nabla I_o(x)|}{|\nabla I_{rb}(x)|} = \sqrt{\frac{\sigma^2+\sigma_0^2}{\sigma^2}} \tag{5.30}$$

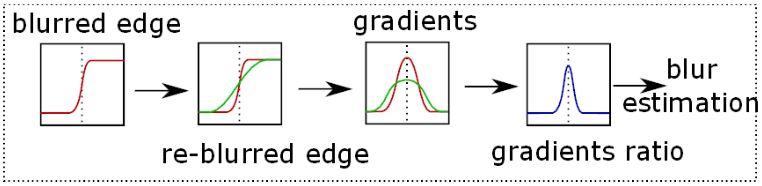


Figure 5.9: Gradient ratio used to measure blur in pictures

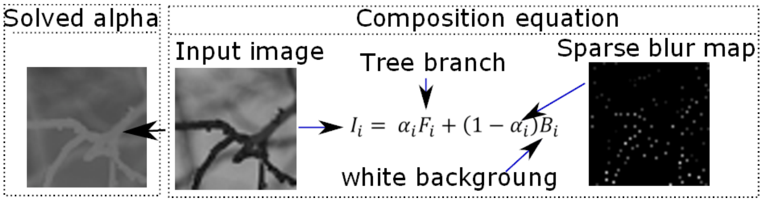


Figure 5.10: Composition equation used to interpolate blur values

$$\sigma = \frac{1}{\sqrt{R^2 - 1}}\sigma_0 \tag{5.31}$$

The second step of the estimation of the dense blur map is to interpolate blur values between the known points. To perform this interpolation, the problem is expressed as a composition equation (eq. 5.32). The image $I_o$ is expressed as the composition of a foreground image and a background image. The value $\alpha$ provide the contribution of the foreground and background image to produce the overall picture under observation. To apply such model to the problem of sparse to dense blur map, the blur measurements obtained previously are considered as ground truth values of $\alpha$. Both images, $F$ and $B$ are unknown. Since the problem is ill-posed, it is needed to define some assumptions about the images. The local smoothness constrains is then assumed in the background picture. Therefore, sharp variation of pixel intensity in picture $I_o$ can only be explained by the foreground image $F$. The objective of such an approach is to obtain the values $\hat{\alpha}$ which will enable to best match to blur measurements and fulfill the smoothness constrains.

$$I_o = \alpha F + (1-\alpha)B \tag{5.32}$$

It can be shown that the solution to this problem can be obtained by solving the sparse linear system of eq. 5.33. The matrix $L$, being the matting Laplacian matrix, and $D$ a diagonal matrix where $D_{i,i}$ is equal to 1 if the pixel $i$ is an edge, and 0 otherwise [142].

$$(L + \lambda D)\,\alpha = \lambda D\hat{\alpha} \tag{5.33}$$

#### 5.3.2.2 Shape from texture

The texture gradient depth cue is estimated in two steps. Similarly to the blur from defocus, a depth map is estimated from the texture gradient, and then a global index is determined. An in-depth description of estimation of the depth map from the texture gradient is can be found in [143]. The main idea of the algorithm is to integrate the gradient field to get the surface which is described by the gradient field. Let $S(x,y)$ be the 2D surface which is expected to be estimated, it is defined on a rectangular grid $\{x = 0,...,W-1; y = 0,...,H-1\}$. Let $p^0 = \frac{\partial S}{\partial x}, q^0 = \frac{\partial S}{\partial y}$ be the integrable gradient field of $S$. The surface $S$, can be exactly recovered by integrating the gradient field $(p^0, q^0)$ by solving a Poisson equation. But with real images the gradient field may not be integrable. Let $(p,q)$ be the non-integrable gradient field and $\hat{S}$ be the estimated surface. Using Simchony, Chellappa and Shao's (SCS) method [144], the surface $\hat{S}$ can be found by minimizing the least square cost function (eq. 5.34).

$$J(\hat{S}) = (\hat{S}_x - p)^2 + (\hat{S}_y - q)^2 \tag{5.34}$$

The Euler-Lagrange equation gives the Poisson equation to solve: $\nabla^2 \hat{S} = div(p,q)$, with *div* the divergence operator, $div(p,q) = \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y} = p_x + p_y$. It could be noted that this equation assumes a null *curl*: $\nabla^2 \hat{S} = div(p,q) = \frac{\partial \hat{S_x}}{\partial x} + \frac{\partial \hat{S_y}}{\partial y} = div(S_x, S_y)$, and then the component $curl(p,q) = \frac{\partial p}{\partial y} - \frac{\partial q}{\partial x}$ is null. The novelty of the approach in [143] is then to take into account the information from the *curl* to increase the accuracy of the surface reconstruction.

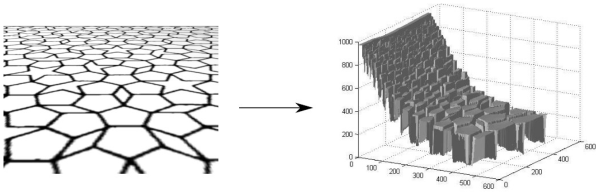


Figure 5.11: Example of result for depth estimation from texture gradient

### 5.3.3 Image segmentation

Image segmentation is a difficult topic which has received a lot of attention, and many advanced techniques have been developed. In the context of this work, it will be useful to decompose the scene into objects composing the scene enabling object-based analysis. To perform this segmentation, it is proposed to use the mean-shift algorithm. The mean-shift is a classical approach for non-parametric clustering which does not require prior knowledge about the number of classes.

For $n$ data points $x_i, i = 1, ..., n$ in a space with a dimension $d$. The algorithm is an iterative process which determine the maximum density of a distribution. For a given initialization, it is possible to determine the kernel density estimate when a kernel K(x) and a window radius h is considered:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5.35}$$

In case of a radial symmetric kernel, e.g. independent of the orientation, the kernel K takes values which fulfill:

$$K(x) = c_{k,d} k(||x||^2) \tag{5.36}$$

With $c_{k,d}$ a normalization constant to ensure that integrating the Kernel provide a value of 1. When this is defined, the goal is to determine the maximum density of the distribution starting from this initialization. To do so, the gradient of the kernel density estimate is computed. It will provide the shift of position compared to the position of initialization, which is called the mean-shift vector.

$$\nabla f(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) g\left(||\frac{x - x_i}{h}||^2\right) \tag{5.37}$$

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g(||\frac{x - x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x - x_i}{h}||^2)} - x \tag{5.38}$$

Based on these equations, the process of the mean-shift is then to:

- Initialize a starting point

- Compute the mean-shift vector $m_h(x^t)$
- Translate the window to a new location: $x^{t+1} = x^t + m_h(x^t)$
- Iterate until the mean-shift vector is null.

During the process, all the points which converge to the same maximum density will belong to the same class.

Using this iterative approach, it was then possible in this work to decompose the scenes into objects. Figure 5.12 depicts some examples of results of the mean-shift applied to the image database used in this work.

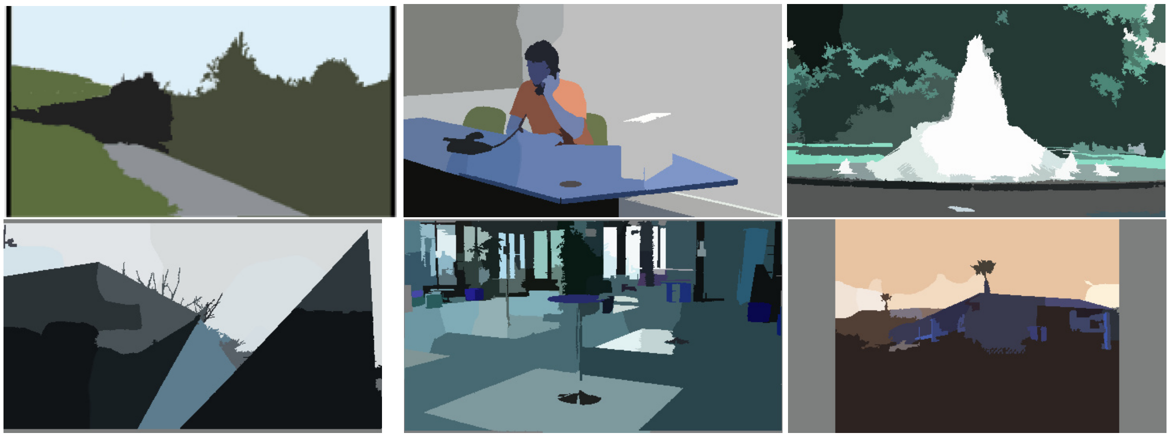


Figure 5.12: Example of image segmentation using the mean-shift.

## 5.4 Binocular depth cues

In previous chapters, the question of evaluating the properties of 3D video sequences was addressed. Until now, subjective methods were used to study the properties of the 3D video sequences. This section will address how the evaluation of binocular depth cues can be performed by means of prediction algorithms. The general structure of the model is depicted in Figure 5.13. There are four main steps: 1) Extraction of disparity maps, 2) identification of regions of depth-interest, 3) feature extraction from selected areas, and 4) pooling of features to calculate the final depth score.

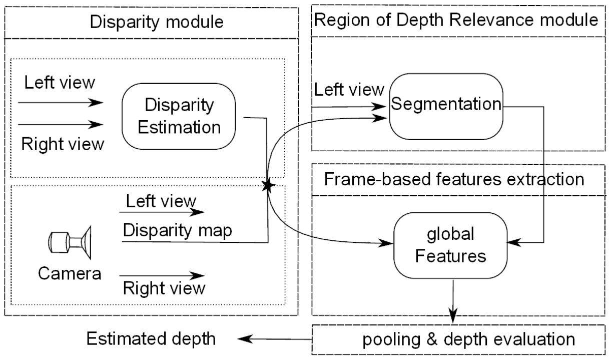


Figure 5.13: General structure of the proposed depth model

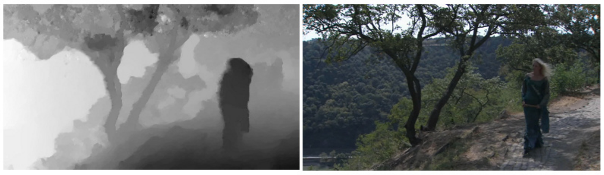


Figure 5.14: Results for the estimation of the disparity

### 5.4.1 *Disparity module*

The target of this first module is to extract a disparity representation that captures the binocular cues and particularly binocular disparities. The most accurate way is to acquire disparity information from the video camera during shooting. Indeed some video cameras are equipped with sensors which provide the ability to record the depth. Using these depth maps, disparities can easily be obtained. At present, it is still rare to have video sequences including their respective depth map. In the future this will be more frequent due to the use of video plus depth-based coding, which will be applied to efficiently encode multiple views as required, for example, for the next generation of multiview autostereoscopic displays. For the present study, this information was not available, and has to be estimated from the two views. To estimate depth maps there exists the Depth Estimation Reference Software (DERS) [145] used by MPEG. This software can provide precise disparity maps. However, it requires at least 3 different views, and information about the shooting conditions (position & orientation of the cameras, focal distances...), information not available for the present research and employed stereoscopic sequences.

Therefore, as discussed in Section 5.3.1, it has then been decided to use a dense optical flow algorithm to estimate the dense disparity maps. An extensive comparison of dense optical flow algorithms is reported by the University of Middlebury [146]. Based on these results the algorithm proposed by Werlberger et al. [136] [99] and available at GPU4Vision [147] was used to estimate disparities from stereoscopic views since it is ranked between the algorithm which provides the best performance and is also particularly fast. This motion estimation is based on low-level image segmentation to tackle the problem of poorly textured regions, occlusions and small-scale image structures. It was applied to find the “displacement” between the left and right stereoscopic views, providing an approximation of the disparity maps. The results obtained are quite accurate as illustrated in Figure 5.14, and are obtained in a reasonable computation time (less than a second for processing a pair of full HD frames on an NVidia GTX470).

Once pixel disparities were computed, it is necessary to convert them to retinal disparities. Figure 5.15 and 5.16 depicts the geometric relationship between the different factors involved in the computation of retinal disparities. Cormack

[148], provided the description of the different equations. If a person is looking at a specific point (F) in space, the angle $cf$ in Figure 5.15 can be computed using the equations 5.39-5.41.

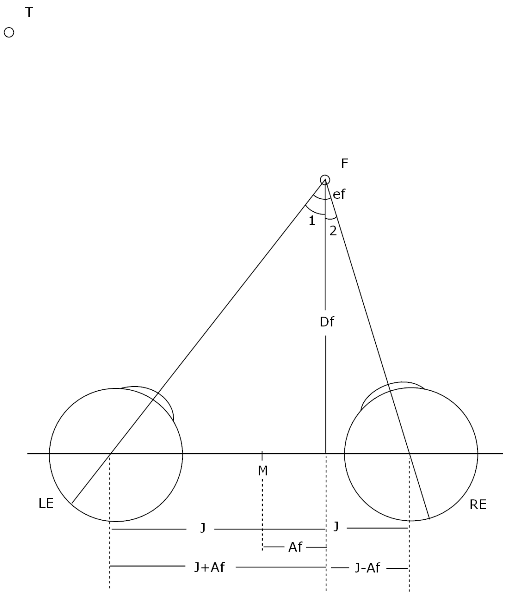


Figure 5.15: Geometric relationship when the eyes fixate a point (F) in space. Figure copied from [148].

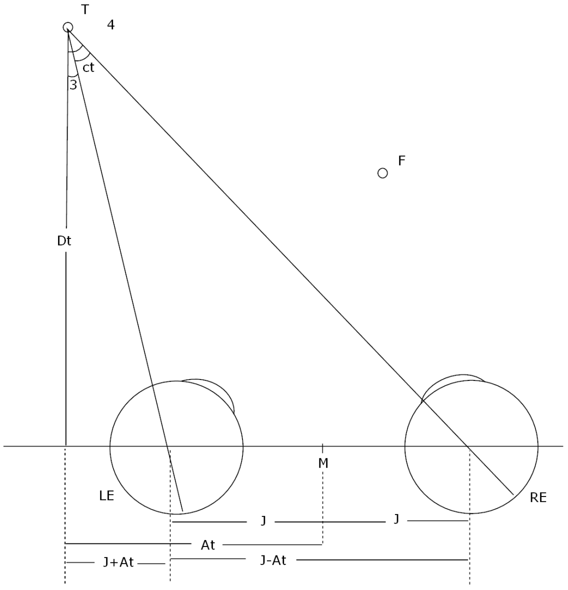


Figure 5.16: Geometric relationship when the eyes are stimulated by a tangent (T). Figure copied from [148].

$$angle\ cf = angle1 + angle2 \tag{5.39}$$

$$angle1 = tan^{-1}\left(\frac{J+Af}{Df}\right) \tag{5.40}$$

$$angle2 = tan^{-1}\left(\frac{J-Af}{Df}\right) \tag{5.41}$$

Then, if a second point is considered, it is similarly possible to determine the angle $ct$ in Figure 5.16 using equations 5.42-5.44.

$$angle\ ct = angle3 + angle4 \tag{5.42}$$

$$angle3 = tan^{-1}\left(\frac{J+At}{Dt}\right) \tag{5.43}$$

$$angle4 = tan^{-1}\left(\frac{J-At}{Dt}\right) \tag{5.44}$$

Finally, the retinal disparity can be computed by performing the difference between $cf$ and $ct$ (Eq. 5.46)

$$\begin{align} r &= angle\ cf + angle\ ct \tag{5.45} \\ &= \left(tan^{-1}\left(\frac{J+Af}{Df}\right) + tan^{-1}\left(\frac{J-Af}{Df}\right)\right) - \left(tan^{-1}\left(\frac{J+At}{Dt}\right) + tan^{-1}\left(\frac{J-At}{Dt}\right)\right) \tag{5.46} \end{align}$$

A particular case of the equation 5.46 is when the convergence is symmetrical and both points $F$ and $T$ are on the midsaggital plane. In this case, $Af$ and $At$ are null. And simplify the equation of retinal disparity (Eq. 5.47)

$$r = 2 \cdot tan^{-1}\left(\frac{J}{Df}\right) - 2 \cdot tan^{-1}\left(\frac{J}{Dt}\right) \tag{5.47}$$

Computing retinal disparities may, however, be challenging since they are the comparison between two distinct points. Processing a $N \times M$ image with a width of $M$ and $M$ the respective height and width of the image would result in $(N \times M)^2$ points since each point need to be compared to each other. Alternatively to keep a perceptual representation taking into account the viewing condition, it is proposed to work with parallax values in degree. Lin [149], provided the equation to compute parallax values, and is described in equations 5.48 - 5.50

$$P = a - b \tag{5.48}$$

$$a = tan^{-1}\left(\frac{D_{IP} + D_s - 2T_s}{2L}\right) + tan^{-1}\left(\frac{D_{IP} + D_s + 2T_s}{2L}\right) \tag{5.49}$$

$$b = tan^{-1}\left(\frac{D_{IP} + D_s}{2L}\right) + tan^{-1}\left(\frac{D_{IP} + D_s}{2L}\right) \tag{5.50}$$

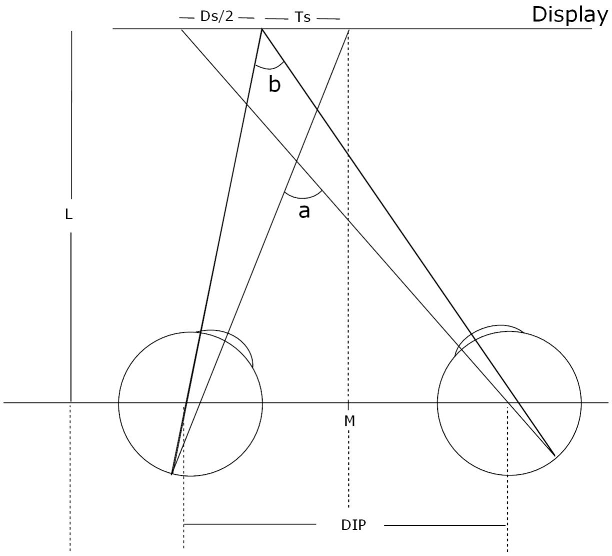
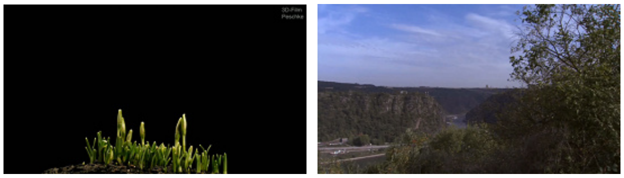


Figure 5.17: Geometric relationship for the computation of parallax.

After processing the disparity map in pixels to convert it to parallax in degree, the next module of the algorithm can be applied.

### 5.4.2 Region of depth relevance module

The idea of the region of depth relevance module is that observers are assumed to judge the depth of a 3D image using areas or objects which will attract their attention and not necessarily on the entire picture, because during scene analysis the combination of depth cues seems to lead to an object-related figure-ground segregation. For example, for the sequence depicted in Figure 5.18a, people are assumed to appreciate the spatial rendition of the grass, and base their rating on it without considering the black background. In the same way, for the scene shown in Figure 5.18b observers are expected to perceive an appreciable depth effect, due to the spatial rendition of the trees and in spite of most of the remaining elements of the scene being flat. Note that this is due to the shooting conditions. The background objects are far away, and hence the depth resolution is low, so that all objects appear at a constant disparity. Further note that the disparity feature provides mainly relative depth information, but it can also give some absolute depth information if the vergence cues are also considered. The region of depth relevance module extracts the areas of the image where the disparities changes, and this way contribute as a relevant depth cue. It is most likely that these areas will be used to judge the depth of the scene. In practice, the proposed algorithm follows the process described in listing 5.1 (also depicted in Figure 5.19):



(a) (b)

Figure 5.18: Illustration of cases where it is assumed that not the entire image is used for judging the depth.

**Listing 5.1: Estimation of the region of depth relevance**

```
---------------------------------------------
Let the function $Std$, the standard deviation as defined by:

$Std : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$
    $X \to \sqrt{\frac{1}{\#X} \sum_{i=1}^{\#X} (X_i - \bar{X})^2}$

With $\bar{X}$ the average value of the elements in $X$
And $\#X$ the cardinal of $X$

---------------------------------------------
Let the variables:

$M$, $N$, $T$: Respectively the number of lines, the number rows of the images and the number
     of frames in the sequence.

$LeftView = [I^L_{n,i,j}]_{N \times M \times T}$,
          $\forall (i,j,n) \in [1,N] \times [1,M] \times [1,T], I^L_{n,i,j} \in [0,255]^3$

$I^L_{n,i,j}$: The pixel value of the left stereoscopic view at the location $(i,j)$ of the frame $n$


$RightView = [I^R_{n,i,j}]_{N \times M \times T}$,
          $\forall (i,j,n) \in [1,N] \times [1,M] \times [1,T], I^R_{n,i,j} \in [0,255]^3$

$I^R_{n,i,j}$: The pixel value of the right stereoscopic view at the location $(i,j)$ of the frame $n$


$Disparity = [D_{n,i,j}]_{N \times M \times T}$,
              $\forall (i,j,n) \in [1,N] \times [1,M] \times [1,T], D_{n,i,j} \in \mathbb{R}$
```

$D_{n,i,j}$: The horizontal displacement of the pixel $I^R_{n,i,j}$ compared to $I^L_{n,i,j}$ such that $I^L_{n,i,j+D_{n,i,j}} = I^R_{n,i,j}$. Here, $D_{n,i,j}$ is the output of the disparity module described in Section 5.A.

$Labels = [L_{n,i,j}]_{N \times M \times T},$
$\forall (i,j,n) \in [1,N] \times [1,M] \times [1,T], L_{n,i,j} \in \mathbb{N}$

$L_{n,i,j}$: The value of the label at the location $(i,j)$ of the frame $n$ resulting of the object segmentation of the left frame using the mean-shift algorithm.

--------------------------------------------

Let *region of depth relevance* as defined by:

For each object, determine the standard deviation of disparity values within the object

$V = [v_{n,l}]_{T \times \mathbb{N}}, \forall (n,l) \in [1,T] \times \mathbb{N}, v_{n,l} \in \mathbb{R}$

$\forall l \in [1, max(Labels)], v_{n,l} = Std(D_{n,i,j})$
$,(i,j) \in [1,M] \times [1,N], L_{n,i,j} = l$

The *region of depth relevance* of the frame n $rodr_n$ is the union of the objects which have a standard deviation of disparity value greater than $dth$

$RODR = [rodr_n]_T, \forall n \in [1,T], rodr_n \in ([1,N] \times [1,M])^{\mathbb{N}}$

$rodr_n = \{(i,j) | (i,j) \in [1,M] \times [1,N],$
$\exists l \in [1, max(Labels_n)] | L_{n,i,j} = l, v_{n,l} > dth\}$

In our implementation $dth$ is set to $0.04$

In the description of the *region of depth relevance* extraction, the mean shift algorithm has been used [150] [151]. This algorithm was discussed in Section 5.3.3 of this thesis, and as previously explained it has been chosen due to its good performance in object segmentation on the data base under study, which has been verified qualitatively for the segmented objects of a random selection of scenes.

### 5.4.3 Frame-based feature extraction module

Once RODR per frame extracted, the next step is to extract the binocular feature used for depth estimation for the entire sequence. The disparities contribute to the depth perception in a relative manner, which is why the variation of disparities between the different objects of the scene are used by the proposed algorithm for depth estimation. In practice, the proposed algorithm follows the lines described in listing 5.2, as illustrated in Figure 5.20:

**Listing 5.2: Estraction of feature per frames**

The frame-based indicator is the logarithm of the standard deviation of the disparity values within the RODR normalized by the surface of the RODR.

$SD = [Sd_n]_T, \forall n \in [1,T], Sd_n \in \mathbb{R}^{\mathbb{N}}$

$Sd_n = \{D_{n,i,j} | (i,j) \in rodr_n\}$

$FrameBasedIndicator = [FrameBasedIndicator_n]_T,$
$\forall n \in [1,T], FrameBasedIndicator_n \in \mathbb{R}$

$$FrameBasedIndicator_n = Log(\frac{Std(Sd_n)}{\#Sd_n})$$

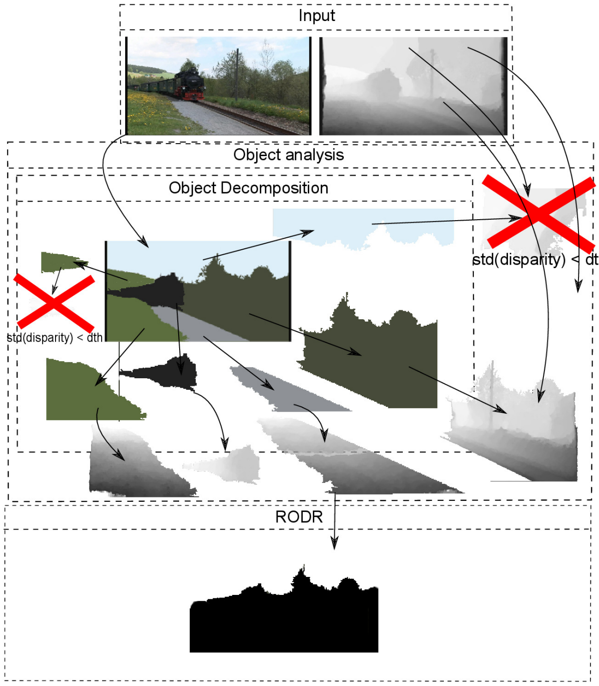


Figure 5.19: Illustration of the algorithm used for determining the region of depth relevance (RODR).

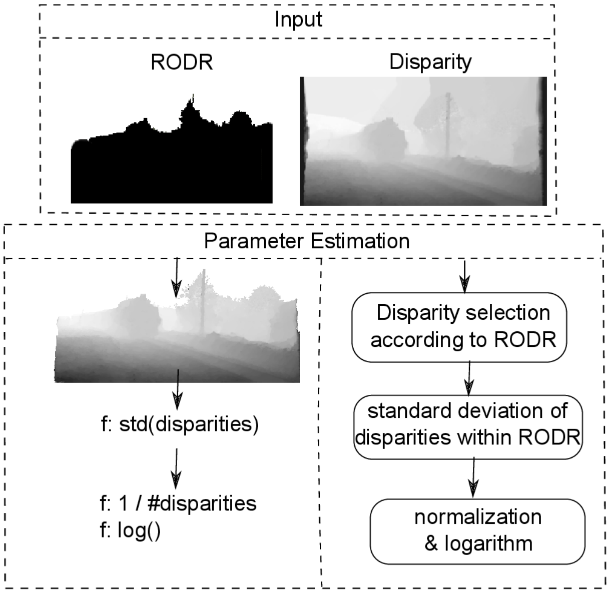

Input
RODR
Disparity
Parameter Estimation
f: std(disparities)
f: 1 / #disparities
f: log()
Disparity selection according to RODR
standard deviation of disparities within RODR
normalization & logarithm

Figure 5.20: Algorithm used for determining the value of the depth indicator for a single frame

### *5.4.4 Temporal pooling*

No temporal properties of the 3D video sequences have been considered so far. To extend the application of our approach from images to the entire video sequences as they are under study in this work, the integration to an overall depth score has to be taken into account. Two main temporal scales can be considered, a local and a global one.

#### 5.4.4.1 Short-term spatio-Temporal indicator

Locally, the temporal depth variation can be used as a reference to understand the relative position of the elements of the scenes. In the previous step, the evaluation of the relative variations in depth of objects per image have been considered, which are extended to a small number of subsequent images to address short term memory, since depth perception is expected to rely on the comparison between objects for consecutive frames. Since the fixation time is 200ms [152], it has been decided to take the temporal neighbourhood into account by analyzing the local temporal variation of relative depth between objects for the evaluation of every frame, to reflect the temporal variation used for evaluating the current frame. A sliding window of *LT* frames corresponding to the fixation time and centered on the frame under consideration was used for the spatio-temporal extension of the depth indicator. In practice, the algorithm is as implemented in listing 5.3, and illustrated in Figure 5.21:

**Listing 5.3: Spatio-Temporal depth indicator**

---------------------------------------------

Let the variables:

$L^T \in 2\mathbb{N}+1$ the size of a local temporal pooling window (for a frame rate of 25 frames per second, $L^T = 5$)

$STdisp = [stdisp_n]_{T-L^T-1}, \forall n \in [1, T-L^T-1], stdisp_n \in \mathbb{R}^{\mathbb{N}}$

$stdisp_n$ the spatio-temporal disparities used for depth evaluation of frame n as in Section 5.C / Listing 2..

---------------------------------------------

The spatio-temporal indicator is defined as:

$stdisp_n = \{D_{t,i,j} | t \in [n - \frac{L^T-1}{2}, n + \frac{L^T-1}{2}], (i,j) \in rodr_n\}$

$STIndicator = [STIndicator_n]_{T-L^T-1}, \forall n \in [1, T-L^T-1], stdisp_n \in \mathbb{R}$

$STIndicator_n = Log(\frac{std(stdisp_n)}{\#stdisp_n})$

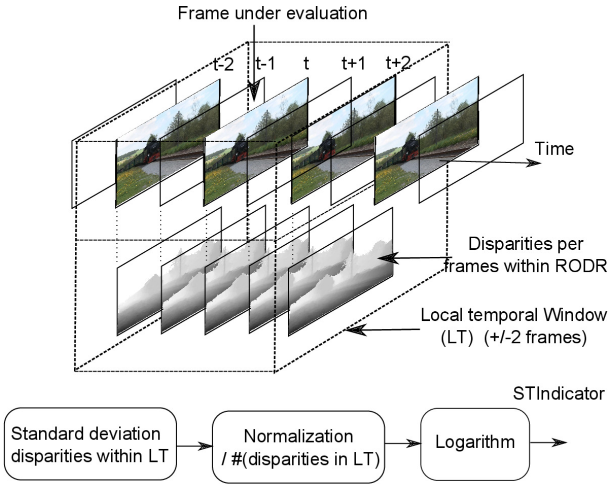

Frame under evaluation
t-2
t-1
t
t+1
t+2
Time
Disparities per frames within RODR
Local temporal Window (LT) (+/-2 frames)
STIndicator
Standard deviation disparities within LT
Normalization / #(disparities in LT)
Logarithm

Figure 5.21: Local temporal pooling

#### 5.4.4.2 Global temporal pooling

Global temporal pooling still require work: it is not trivial to pool the different instantaneous measures to calculate an estimate of the global judgment as obtained from the observer. In the case of quality assessment, there are several approaches for temporal pooling, such as the very simple averaging, Minkovsky summations, average calculation using Ln norm or limited to a certain percentile. Other approaches are more sophisticated [153] and deal with quality degradation events. Regarding the global estimation from several local observations, it is usually assumed that if an error occurs people will quickly say that the overall quality of the sequence is poor, and it will take some time after the last error event until the overall quality is considered as good again [153]. In the context of our depth evaluation, this seems to be the inverse: observers who clearly perceived the depth effect will quickly report it, and if there are some passages in the sequence where the depth effect is not too visible, they seem to take some time to report this in their on overall rating. To reflect this consideration on our model, we then decided to use a Minkovsky summation with an order higher than 1, to emphasize passages of high short-term depth-values. The final mapping is then performed using a third order polynomial function.

**Listing 5.4: Global temporal pooling**

$$Indicator = \frac{1}{T-L^T-1}\sqrt[k]{\sum_{t=1+\frac{L^T}{2}}^{T-\frac{L^T}{2}} (STIndicator_t)^k}$$

In our implementation $k$ is set to 4

$$MOS_e = A \times Indicator^3 + B \times Indicator^2 + C \times Indicator + D$$

In our implementation $A$, $B$, $C$, $D$ are respectively set to $-0.06064$, $-2.213$, $-25.79$, $-93.04$ (
    obtained by using the optimization function polyfit of MATLAB)

## 5.5 Model performance

To evaluate the performance of the proposed model, the subjective video database created and described in section 4.4 was used to measure the accuracy of the algorithm. Figure 5.22 depicts the subjective depth ratings as compared to the the predicted subjective depth. The model training and validation are carried out using cross - validation (6 combinations of training/validation). The model achieves the following performance: the Pearson correlation R = 0.60, the root mean squared error RMSE = 0.55, the RMSE* = 0.37, and the outlier ratio (OR) is equal to 0.83 / 21.33 (where 0.83 is the number of outliers on a validation dataset subset composed of 21.33 sequences. The reported floating point values are mean values which stem from the cross-validation) on our entire database for seven defined parameters (The threshold in the RODR algorithm,the size of the local temporal pooling, the order of the Minkovsky summation, the four coefficients of the polynomial mapping). These results show that there is still space for further improvements.
As it can be observed from Figure 5.22, eight source sequences are not well considered by the algorithm (plotted as red triangles). These specific contents show a pop-out effect which apparently was well appreciated by the observers, who rated these sequences with high depth scores. Two distinct reasons could explain these results: From a conceptual point of view, the current algorithm does not make a difference between positive and negative disparity values, and hence between the cases that the objects pop out or stay inside the screen. From an implementation point of view, the disparity algorithm did not succeed to well capture the small blurry objects that characterize the pop-out effect. This leads to an under-estimation of the depth for these contents. Without these contents, we achieve a Pearson correlation of 0.8, an RMSE of 0.38, an RMSE* of 0.18 and an OR of 0 / 18.66.
Even though they do not have a strong effect on the general results of the model, a second type of contents could also be identified (represented by the circles in Figure 5.22), which have been overestimated in terms of depth. For the lower contents, it is still unclear what factors contribute to these ratings. Some of the sequences show fast motion, some have several scene changes, and other depth cues may also inhibit the depth perception.

As a consequence, three factors are currently under study to improve the general accuracy of the model:

- Incorporate a weighting depending on the position in depth of the object (if they pop-out or stay inside the display),
- Improve the accuracy of the disparity estimation,
- Consider the monocular cues which are in conflict with the binocular depth perception.

**Listing 5.5: RMSE***

```
let $X_{gth} \in \mathbb{R}^{\mathbb{N}}$ a set containing the ground truth values.
let $X_{est} \in \mathbb{R}^{\mathbb{N}}$ a set containing the estimated values.

let $CI_i^{95}$ the confidence interval at 95% of $X_{gth_i}$

$\forall i \in [1, \#X], P_{error_i} = max(0, |X_{gth_i} - X_{est_i}| - CI_i^{95})$

$RMSE^* : \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$
        $(X_{gth}, X_{est}) \rightarrow \sqrt{\frac{1}{\#X - d} \sum_{i=1}^{\#X} (P_{error_i})^2}$

With $d$ the degree of freedom between $X_{gth}$ and $X_{est}$
```

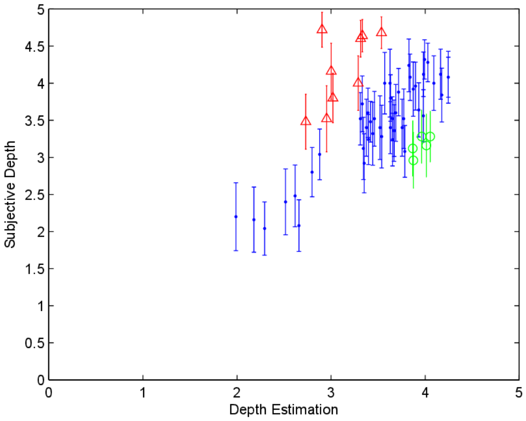


Figure 5.22: Results of the model on the estimation of depth, the triangles represents the contents which have pop-out effect, the circle represents a class of under-estimated content (which have a lot of linear perspective)

## 5.6 Applications

One of the main issues in the subjective evaluations of the perceived depth as a function of monocular and binocular depth cues in natural images as presented in the section 4.5.3 using absolute category rating and section 4.6.1 using ranking, is the differentiation between the contribution of binoculars and monocular cues to the overall depth perception. The uses of natural images in the test have affected the ability to define precisely the amount of monocular and binocular depth cues. Some of the monocular depth cues were evaluated in the images as described in the previous section. The contribution of binocular depth cues in these images is even more difficult to evaluate since it is not possible to consider them independently of the monocular cues. Considering how important binocular depth cues are, different measurements are applied to the images' depth map to provide more detailed information on the perceived depth coming from retinal disparities or from monocular depth cues.

To perform such analysis, it is possible to use the model previously described in this chapter which only takes into account the characterization of the binocular properties of the stereoscopic properties of the images.

**Listing 5.6: Research questions**

```
1. How the proposed algorithm compares with other approaches from the literature.
2. What are the interactions between binocular depth and monocular depth cues.
```

### *5.6.1 Comparison with other methods*

There are only few studies which have addressed computational models for predicting depth perception in natural images, and provide an overall score. Nevertheless, in addition to the proposed method, other alternatives have been proposed in recent years.

Sohn [154], characterized the depth of the images using an object decomposition and applied two different metrics on objects: the mean disparity value and measured the "object thickness" ($OT_s$) by measuring the ratio between the mean width of the object $s$, noted $MW_s$, and mean disparity of the object, noted $MAS_s$. Even though the metric was applied to estimate visual discomfort, it provides another way of depth characterization which can be applied to this study. Extending this approach, Toyosawa [155], added to the work of Sohn [154] the consideration of the depth range between the farthest and closest object into the overall equation.

$$OT_s = ln(\alpha \cdot \frac{MW_s}{MAD_s}) \tag{5.51}$$

Addressing the same use case as studied in this thesis, Toyosawa [155], provided a comparison of 20 different statistical analysis of disparities. It covers many kinds of measurements such as average, minimum, maximum, standard deviation of disparities, different quartile, and disparity range size [154, 156]. Lin [149], proposed a measurement algorithm based second order polynomial fitting of the range of depth values and the average was also considered. It is described by equation 5.53. The notation $R_{n\%}$ meaning the $n^{th}$ percentile of the retinal disparities values.

$$R_{range} = R_{90\%} - R_{10\%} \tag{5.52}$$

$$S^* = a \cdot R_{50\%} + b \cdot R_{range} + c \cdot R_{50\%}^2 + d \cdot R_{range}^2 + e \tag{5.53}$$

The result of the different measurements considered in [155], shows that not only one of them can fit for all cases and the most appropriate measurement algorithm depends on the composition of the image and the number of objects at different depth plane. Unfortunately most of the previous studies only considered a small number of images: 10 different images composed of only one 3D rendered objects in [149], 14 natural images from different movies in [155], 40 natural images in [154] which begins to be reasonably large but the focus of the paper was not perceived depth but visual comfort, and 64 videos in our previous work [113]. To extend the study proposed in [155] it is proposed to add more contents, e.g. the 200 3D still images previously presented in section 4.5.2, and to consider a new high-level measurement algorithm. It is important to note that this is then a different one than the database used for training and previous verification of the model designed in the context of this thesis. The results of the performance of these measurements can be found in Table 5.2 and Figure 5.24. The $RMSE^*$ is the root mean square error (RMSE) taking into account the size of the confidence interval of the subjective data. Let $X_{gth} \in \mathbb{R}^{\mathbb{N}}$ the ground truth values, $X_{est} \in \mathbb{R}^{\mathbb{N}}$ the predicted values, $CI_i^{95}$ the confidence interval at 95% of $X_{gth_i}$, and $d$ the degree of freedom between $X_{gth}$ and $X_{est}$.

$$RMSE^* = \sqrt{\frac{1}{\#X - d} \sum_{i=1}^{\#X} (P_{error_i})^2} \tag{5.54}$$

Table 5.2 depicts the performance of 31 different ways to characterize the 200 images of the studied database. Since each indicator have different ranges than the subjective scores, a third order polynomial fit is applied between the indicators and the respective subjective data enabling to compute outlier ratio (OR), RMSE, RMSE$^*$ (equation 5.54) in addition to the Pearson and Spearman-correlations.

The category "Distance metrics" is composed of different percentile used as an indicator of the content property. The notation PXYz define the value of the percentile $XY.z\%$. Between these indicators, it can be seen that the value of the minimum value of disparity is a better indicator of perceived depth than the maximum values.
In the category "Volume metrics", the indicator $PX_1Y_1z_1 - PX_2Y_2z_2$ determine the size of the interval in disparity between the percentiles $X_1Y_1.z_1\%$ and $X_2Y_2.z_2\%$. The Michelson contrast is defined as in equation 5.55. The "2nd order polynomial fit" correspond to the method described by Lin [149], equation 5.53, and using the coefficient provided into their paper. To provide a better comparison with their algorithm, it is proposed to retrain the coefficient on the proposed database using the MATLAB function "regress". This corresponds to the line "2nd order polynomial refit" in Table 5.2. ITU-T Recommendation P.1401 [157] describes how to perform statistical tests to compare the performance of different prediction algorithms. The Lin algorithm [149] with its original coefficients outperform in terms of Spearman correlation the other volume metrics with the exception of the interval P950-P050. It can be reminded that this algorithm is also based on this interval, which appears to indicate that this parameter is a key feature of the proposed algorithm. However the metrics do not outperform statistically the "distance metrics" based on percentile lower than the median. With regards of the Pearson correlation, the retrained Lin's algorithm provides similar results as the ones described for the Spearman correlation, and achieve a lower RMSE.

$$Contrast = \frac{P950 - P050}{P950 + P050} \tag{5.55}$$

The higher level category "object-based metrics" include the work of Toyosawa [155] with the average thickness of objects, the depth interval between the closest and farther object, the number of objects. The metrics also include the proposed method, and the work of Sohn [154] on objects thickness. The proposed algorithm outperform statistically the other object-based algorithms on each performance indicators. It is statistically equivalent to the Lin [149] volume-based algorithms in terms of Pearson correlation, outperform the retrained Lin algorithm. The Lin algorithm with its original parameter, and the interval $P950 - P050$ outperform our algorithm in terms of Spearman correlation. However, if the RMSE, and RMSE$^*$ is considered, the algorithm [113] provides better performance compared to the other Volume- and Object-based algorithms across the different performance evaluation criteria.
However, compared to the "distance metrics", it can be seen that no algorithm outperform significantly the "simple" 7.5% percentile on each of the performance criteria. Therefore, if a least computing intensive algorithm needs to be defined, this algorithm may be candidate solution even though it lacks of addressing many perceptual aspects, for example having only one depth plane. In this case algorithms such as [149] or [113], may provide a better solution. However, such contents were not available in the studied database.

To conclude, many factors are involved into the overall depth perception and simple statistics about the depth maps appeared not to be able to explain how depth is perceived. The image structure addressed in [113, 154] via object analysis or as the notion of image composition in [155] is one of the directions to address the missing aspects of analysis statistical depth map. This enables tackling aspects such as the monocular depth cues, but a lot of work still

| | Measurements | N° | PC | SC | OR | RMSE | *RMSE*$^*$ |
|---|---|---|---|---|---|---|---|
| Distance metrics | P005 | 1 | 0.192 | 0.498 | 0.016 | 4.188 | 3.487 |
| | P010 | 2 | 0.476 | 0.426 | 0.016 | 3.738 | 2.948 |
| | P015 | 3 | 0.331 | 0.523 | 0.016 | 3.841 | 3.085 |
| | P020 | 4 | 0.474 | 0.525 | 0.011 | 3.774 | 2.982 |
| | P025 | 5 | 0.406 | 0.538 | 0.016 | 3.859 | 3.087 |
| | P050 | 6 | 0.522 | 0.566 | 0.000 | 3.706 | 2.934 |
| | P075 | 7 | 0.526 | 0.582 | 0.000 | 3.707 | 2.936 |
| | P100 | 8 | 0.516 | 0.571 | 0.000 | 3.704 | 2.932 |
| | P125 | 9 | 0.520 | 0.567 | 0.000 | 3.706 | 2.935 |
| | P500 | 10 | 0.449 | 0.448 | 0.021 | 3.686 | 2.910 |
| | P875 | 11 | 0.046 | 0.042 | 0.043 | 4.218 | 3.507 |
| | P900 | 12 | 0.039 | 0.025 | 0.048 | 5.249 | 4.662 |
| | P925 | 13 | 0.034 | 0.106 | 0.048 | 6.970 | 6.509 |
| | P950 | 14 | 0.031 | 0.116 | 0.048 | 7.934 | 7.521 |
| | P975 | 15 | 0.030 | 0.150 | 0.048 | 8.661 | 8.277 |
| | P980 | 16 | 0.030 | 0.178 | 0.048 | 8.670 | 8.287 |
| | P985 | 17 | 0.029 | 0.183 | 0.048 | 8.952 | 8.580 |
| | P990 | 18 | 0.027 | 0.207 | 0.048 | 9.937 | 9.597 |
| | P995 | 19 | 0.025 | 0.266 | 0.043 | 13.358 | 13.088 |
| Volume metrics | P950-P050 | 20 | 0.005 | 0.518 | 0.016 | 25.323 | 25.145 |
| | P975-P025 | 21 | 0.021 | 0.445 | 0.016 | 16.120 | 15.893 |
| | P990-P010 | 22 | -0.004 | 0.402 | 0.027 | 55.227 | 55.111 |
| | standard deviation | 23 | 0.071 | 0.439 | 0.011 | 7.323 | 6.907 |
| | Michelson contrast | 24 | 0.060 | -0.067 | 0.287 | $> 100$ | $> 100$ |
| | 2nd order polynomial fit [149] | 25 | 0.029 | 0.587 | 0.015 | 23.540 | 23.362 |
| | 2nd order polynomial refit [149] | 26 | 0.480 | 0.432 | 0.021 | 4.109 | 3.307 |
| Object metrics | Avg. Thickness [155] | 27 | 0.085 | 0.126 | 0.043 | 3.616 | 2.811 |
| | Depth Interval btw Objects [155] | 28 | 0.054 | 0.276 | 0.038 | 8.191 | 7.810 |
| | Nb Objects | 29 | -0.051 | -0.004 | 0.064 | 3.693 | 2.914 |
| | PerceptualDepthIndicator [113] | 30 | 0.579 | 0.479 | 0.000 | 3.739 | 2.964 |
| | Object thickness [154] | 31 | 0.276 | 0.345 | 0.080 | 3.430 | 2.748 |

Table 5.2: Different algorithms to evaluate binocular depth in 3D contents. PC: Pearson correlation, SC: Spearman Correlation, OR: Outlier Ratio, RMSE: root mean square error, and *RMSE*$^*$ as defined in eq 5.54.

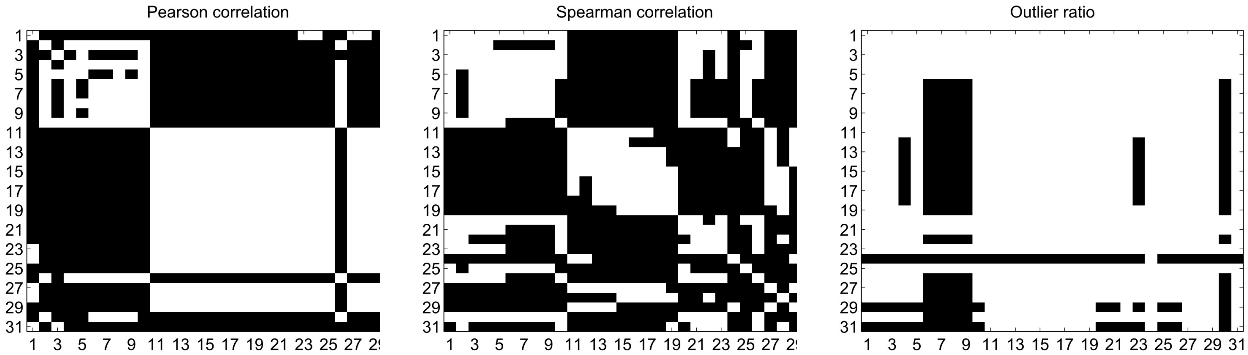


Figure 5.23: Statistical differences between depth indicators. Black indicates statistical differences.

remains to fully characterize them. The next subsection will present the results obtained regarding the relation between perceived depth and monocular depth cues from the subjective experiments conducted in this study.

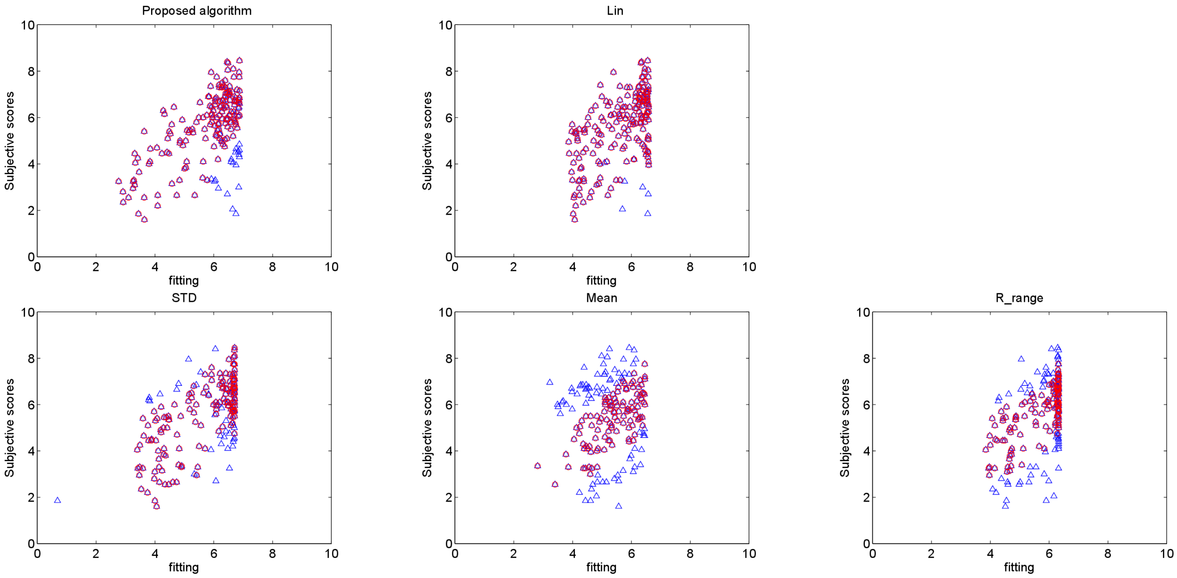


Figure 5.24: Relation between instrumental measurement of image's binocular characteristics and binocular depth scores. Red circles indicates the inliers of the RANSAC fitting.

### 5.6.2 *Depth perception and its relation with monocular and binocular depth cues*

To study how monocular depth cues affect the perceived depth, monocular depth scores are put into relation with binocular depth cues. The distribution of the individual depth cue scores of images is described by the following formula: Let $S_{i,o}$ be the depth score provided by observer $o$, on image $i$ and, $l_i$, $r_i$, $i_i$, $b_i$ be respectively the depth cue scores on the linear perspective, relative size, interposition and binocular scales for this same image $i$. Let $I$ be the set of available images. The probability function is defined by:

$$\forall i \in I, \forall k \in [0,11], \forall f \in l,r,i,b, P(f_i|S=k) = G_{f,k,i} \tag{5.56}$$

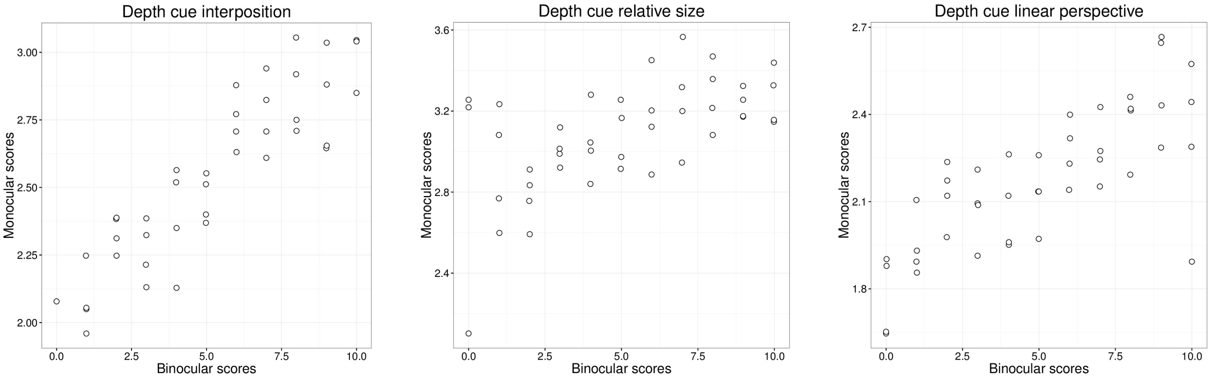


Figure 5.25: Relation between monocular and binocular depth scores

With $G_{f,k,i}(x)$ is the probability of having a score $x$, for the depth cue $f$ for an image $i$, knowing that the overall depth score is $k$. Figure 5.25 depicts the relation between the different factors involved in $G_{f,k,i}(x)$. A Pearson correlation of

0.898, 0.608 and 0.756 can be found between the binocular scores and respectively the interposition, relative size and linear perspective depth cues.

However, based on this data a direct relation between monocular depth cues and overall depth rating cannot be too easily drawn. Indeed, it should be noted that the considered monocular depth cues were only one factor involved in the depth scores, and retinal disparities specifically have a very strong impact on the perceived depth ratings. To study the actual contribution of monocular depth cues, their contributions should be dissociated from the contribution of binocular depth cues. In the case of this study, natural images were used, therefore it is not easily possible to fully characterize what part of the overall depth ratings come from the monocular depth cues, and what comes from the binocular cues. In order to address this issue, the prediction algorithm based on the analysis retinal disparity developed in this thesis and described in Section 5.4 was used to study the contribution of binocular depth cues to the overall depth rating. Figure 5.25 depicts the relationship between the prediction of depth scores using this algorithm and the overall depth scores obtained through subjective evaluation which takes into account both monocular and binocular depth cues. The Pearson correlations between the binocular depth measurements and overall depth scores are respectively: 0.92, 0.676 and 0.795 for the different image sets. These correlation values are found to be in the same range as the respective Pearson correlation values between monocular depth score and overall depth rating for the same set of images. This indicates that the overall depth ratings are also highly related to the retinal disparity distribution.

To further study the relation between the monocular depth ratings and overall perceived depth rating, the monocular depth cues scores are put into relation with the predicted depth scores from the model defined in Section 5.4. A Pearson correlation between the monocular cues “interposition”, “relative size”, “linear perspective” and predicted depth obtained from disparity analysis is computed. The results are found to be respectively: 0.819, 0.606 and 0.663. This shows that the interposition monocular cues score and the binocular measurements are highly related. An explanation is that interposition was subjectively characterized by the fact that multiple layers were visible in the scene. It is expected that the availability of multiple layers relates to the presence of multiple depth layers which affect perceived depth in natural images and explain the high correlation between interposition and the depth scores. Similarly, the binocular depth measurement was defined such as they evaluate the availability of multiple depth layers in the pictures. It is then expected that the interposition depth scores relate to the binocular cues measurement scores.

In the same manner, the availability of linear perspective into the picture with a vanishing point at the center of the picture is also intrinsically related to the presence of different depth layers having lines converging to the vanishing point. This could also explain the higher relationship between linear perspective and both perceived depth and the binocular depth metric.

These results on the linear perspective scales and interposition depth cues appears then more related with intrinsic properties of the image structure. Which by extension relates to the distribution of the binocular depth cues. Therefore, the result of this analysis relates image properties to retinal disparities distribution, than defining a perception model on the combination of monocular depth cues.

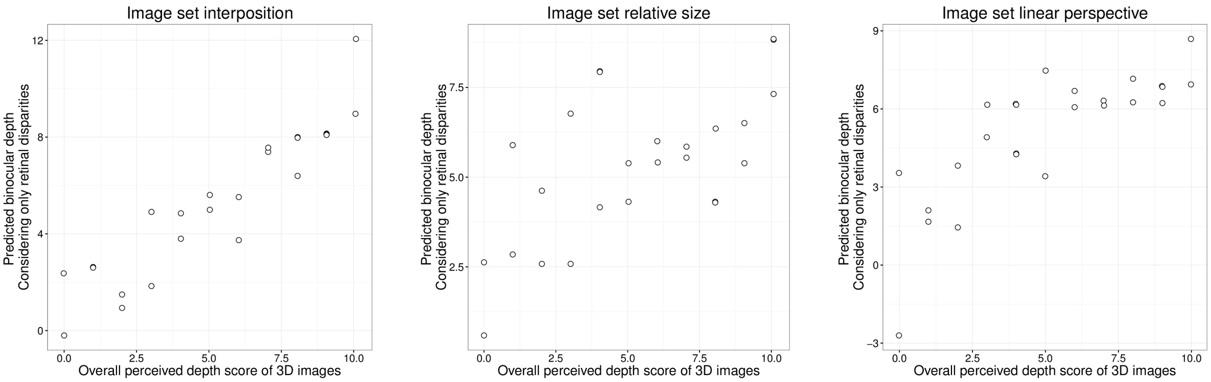


Figure 5.26: Relation between instrumental measurement of image's binocular characteristics and binocular depth scores on the three different set of images designed to study interposition, relative size and linear perspective depth cue.

### 5.6.3 Conclusion

One of the issues addressed in this section is the difficulty to instrumentally characterize the binocular depth cues in natural images. Different kind of measurement from the literature were compared and none of them succeed to fully explain the depth scores. The image structure or image composition, and then monocular depth cues needs are further aspects which can be considered. The relation between monocular and binocular depth scores have shown that monocular characteristics of the images affected the binocular depth scores. However, considering the relation between the monocular scores and the binocular measurements though instrumental measurements, the conclusion is: the relative size and linear perspective depth cues implies multiple depth layers which induces a higher perceived depth. But due to the lack of appropriate binocular depth cue characterization it is not yet possible to study depth cues combination with natural images.

**Listing 5.7: Conclusion on the research questions**

1. Using a simple indicator such as the 7.5 percentile already provides interesting result on evaluating perceived depth, even though it will not be able to address special cases such as images suffering of cardboard effect.
2. The proposed model performed better than the other approaches presented in this work. However, the performances are still not satisfactory and there is space for improvement.
3. Monocular characteristics of the images affected the binocular depth scores. However, it is not clear whether the monocular depth cues themselves affected the depth percept or if the presence of these depth cues implies specific image properties which changes the depth percept.

## 5.7 Monocular depth cues

One important aspect as described in Section 2.2.2, is the contribution of monocular depth cues to the depth perception. In order to tackle the evaluation and characterization of these depth cues different instrumental evaluation methods of depth cues were studied in the context of this work and will be the purpose of this section.

### *5.7.1 Linear perspective*

The linear perspective depth cue was presented in section 2.2.2, and attempt to characterize it in subjective evaluation was done in section 4.3. In this subsection methods for instrumental evaluation will be provided.

#### 5.7.1.1 Subjective evaluation database

In order to develop meaningful instrumental evaluation methods for linear perspective, it is necessary to dispose of the ground truth data which will be used both for training and verification purpose. The database developed by Ross and Oliva [158] was used. It consists of 7138 different images with various types of content: nature or urban scenes (Figure 5.27 provides an illustration of the category of pictures). All the different images were annotated by 14 participants in total. The test participants had to rate the content of the scene on three distinct scales: the perspective, the depth, and the openness. The perspective was defined as following: “Perspective refers to the degree of expansion of space. The convergence of parallel lines to a visible vanishing point gives a strong perception of depth gradient to the space represented in an image” [158]. The depth scale was provided as: “the size of the space in a scene (e.g. the mean distance between the observer and the boundaries of this space, e.g). While dominant depth is not a precisely defined quantity, it has a strong relationship with the physical size of the space, and human judgments are consistent in evaluating this quantity” [158]. The openness was defined by “the quantity and location of boundary elements of the scene in view. The most open scene is a ground surface stretching to the horizon, with the existence of a horizon line in the absence of any other visual references (e.g. trees, buildings).” [158]. Each scale was discrete, and composed of 6 points. Example pictures were provided to the test participants to guide them to rate the images.
Unfortunately, considering how large the task was, participants did not rate all the images but only a subset of the 7138 images. This has resulted in at most two ratings per image which happened to 12% of the images. This is the major problem of this database. The author studied the relationship between the scores of the images rated by two test participants which shows in some cases a large variance between replies. This variance between scores was used as a measure of the precision of the ground truth. In this subsection, only the linear perspective scores will be studied.

#### 5.7.1.2 Candidate evaluation algorithms

Two different approaches were evaluated to predict the presence of linear perspective and its relationship with perceived depth in images.

First approach

The first approach considered is the one proposed by Ross and Oliva [158] and is called the “Global layout properties” (GLP) and is based on the set defined by Oliva for scene description [159]. The images are decomposed into non-overlapping blocs which are then described through the set of GIST features [159]. The blocks are filtered through different Gabor filters having different orientation (8) and frequency bandwidth (4) (See Figure 5.28). The magnitude of each resulting filtered block is then averaged. A principal component analysis is used to reduce the dimension of

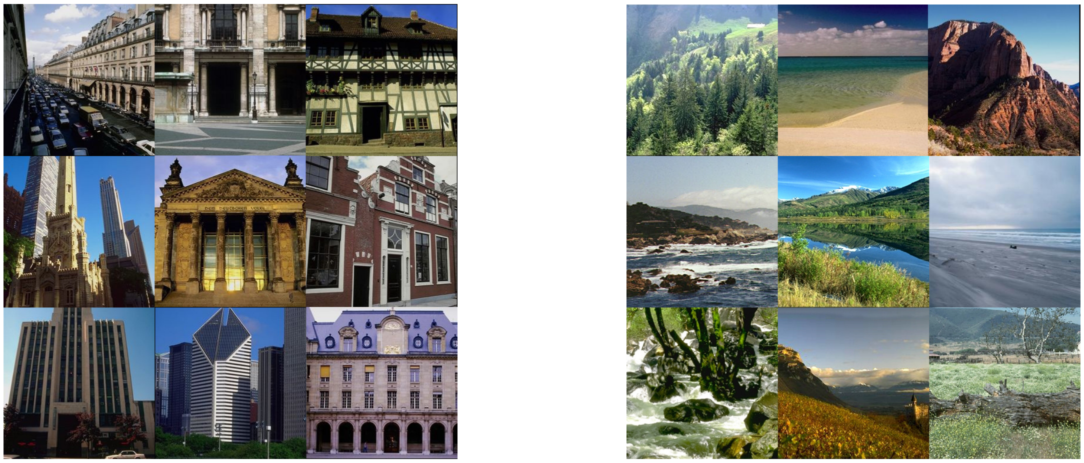


Figure 5.27: Illustration of the image database from Ross et Oliva [158]. Pictures are divided into two categories: urban (left), natural (right).

the features to 24 features, and finally a cluster weighted model is then trained to predict the strength of the linear perspective.

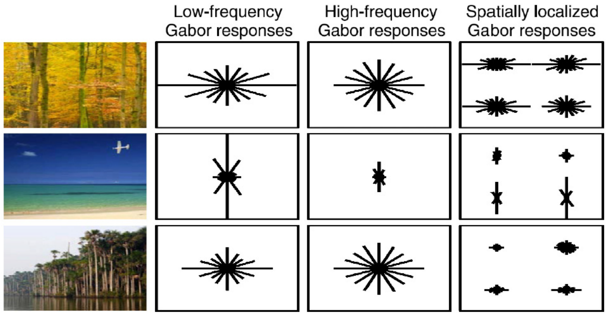

Low-frequency Gabor responses
High-frequency Gabor responses
Spatially localized Gabor responses

Figure 5.28: GIST features: decomposition of blocks into different Gabor filter having different orientation and frequency bandwidth. Figure from [160]

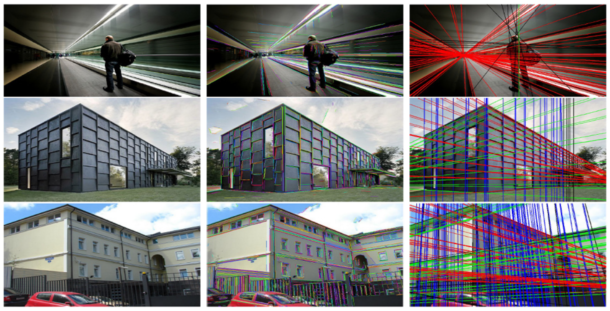


Figure 5.29: Vanish point model applied on different images. Left, input images; middle, line segment; right, vanishing lines

Second approach

The second method employed is called the vanish point model, this method address the geometric properties of the scene as depicted in Figure 5.29. A line segment detector (LSD [161]) is employed to extract lines. The vanishing points are determined by using the J-Linkage algorithm to determine which lines converges to the same vanishing point. As a final step, it was proposed to define the strength of the linear perspective depth cue as a function of the distance of the vanishing point to the center of the image (See eq. 5.57), with $d$ the distance of the closest vanishing point to the center of the image (Figure 5.30). This rule was derived from the observation of the images available in [158] and is depicted in Figure 5.31.

$$x = \frac{1}{d+1} \tag{5.57}$$

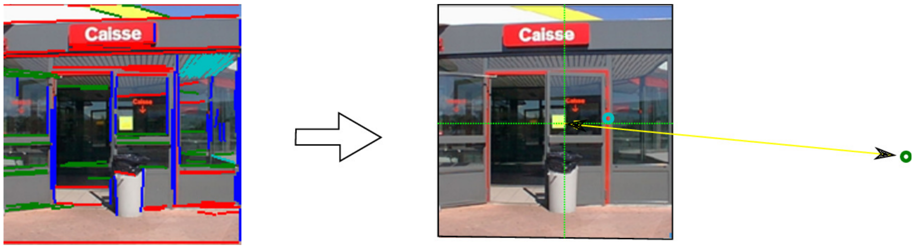


Figure 5.30: General principle of the vanish point model.

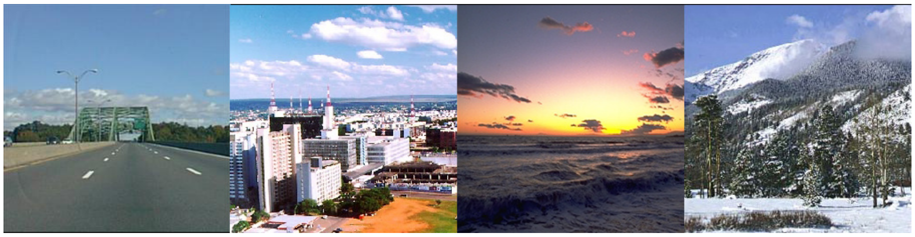


Figure 5.31: Empiric rule for linear perspective. Images ordered from strongest linear perspective to lowest.

#### 5.7.1.3 Results

The performance of the proposed methods was evaluated by measuring their ability to predict the scores obtained by the test participants. As mentioned previously, there exists two different kinds of images to be predicted: the "urban" and "natural" scenes. Considering the difference of performance of the metrics for each class of images, it is proposed to present the result per group of images. As depicted in Figures 5.32 - 5.35, the performance of both approaches are better in case of urban than natural scenes (Pearson correlation of 0.64 and 0.59 instead of 0.33 and 0.17). In the latter case, the scatter plots are close to random.
The metric limits were further studied by analyzing the outliers. It revealed that GLP usually underestimates the linear perspective in case of strong texture distributed over the entire picture. On the contrary, the vanish point model will underestimate the linear perspective in case of images having only few vanishing lines. A more in-depth analysis of the limits of metric's performance will be provided in the Section 5.8.2 where the question of metric reliability and identification of failure cases is addressed.

### *5.7.2 Defocus blur*

The second monocular depth cue which was investigated is the defocus blur. To characterize this cue, first a dense blur map was computed as described in the Section 5.3.2.1. Then, it is necessary to translate the blur measurements to depth values. Finally, the depth indicator can be computed. The process of measuring blur in images was described extensively in Section 5.3.2.1, therefore in this section the focus will be on the conversion to depth, measurement of algorithm performance and depth indicator computation.

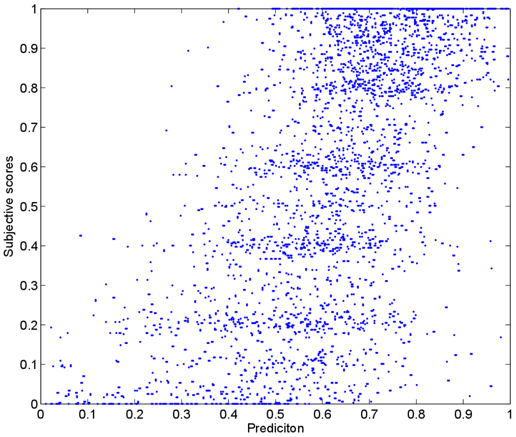


Figure 5.32: GLP performance on urban scenes

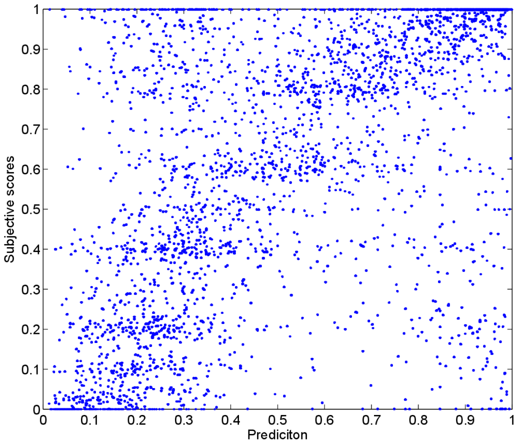


Figure 5.33: VPM performance on urban scenes

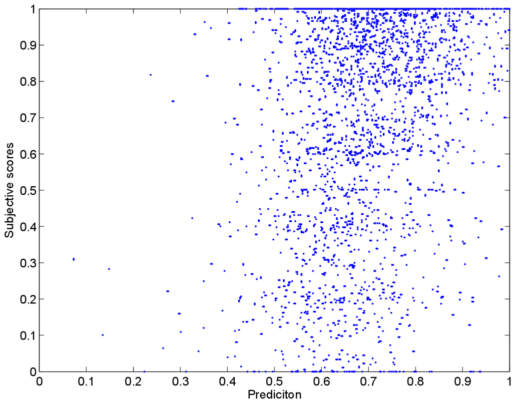


Figure 5.34: GLP performance on natural scenes

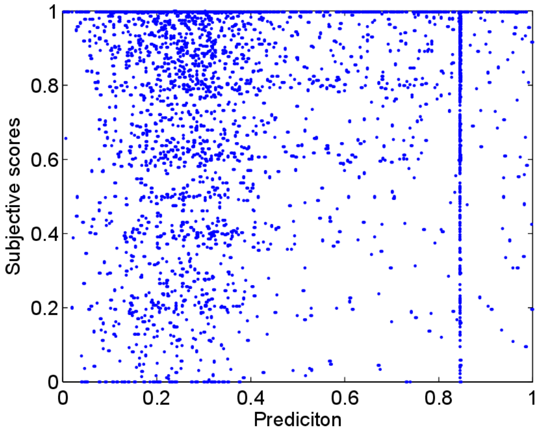


Figure 5.35: VPM performance on natural scenes

#### 5.7.2.1 Conversion of blur to depth

The circle of confusion is the optical spot from the intersection of the cone described by the light rays from the lens and the optimal focus point (see Figure 5.36). Its size depends on several parameters: the distance between the object and the lens $d$, the distance between the lens and the focal plane $d_f$, the focal length $f_0$, the aperture $N$ defined relatively to the focal length. Based on these parameters the circle of confusion, $c$, can be determined as in Eq. (5.58).

$$c = \frac{|d - d_f|}{d} \frac{f_0^2}{N(d_f - f_0)} \tag{5.58}$$

To estimate the distance between the blurred object and the lens it is possible to reverse Eq 5.58, by taking into account that $d \geq d_f$, and $d - d_f \geq 0$ it is possible to remove the absolute values, and $d$ can be expressed as in Eq. 5.59. In this equation it was necessary to introduce a parameter $k$ such as $\sigma = kc$ which take into account the resolution and size of the sensor noted respectively $Sensor_W$ and $Res_W$. Within the parameter $k$, the amplitude of the Gaussian blur is also included. This finally enables to convert the blur measured in pixel in the previous section to a circle of confusion in meter.

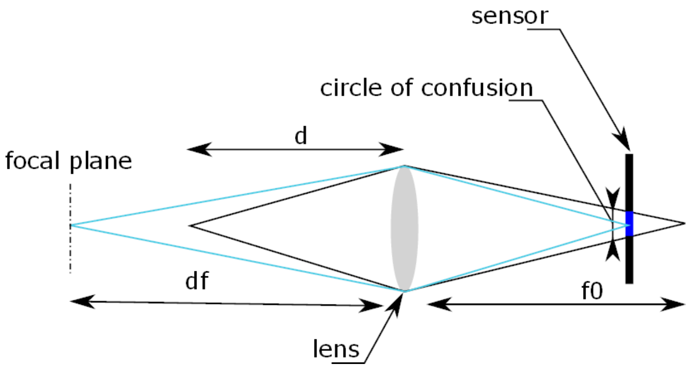


Figure 5.36: Circle of confusion and relation with capture settings.

$$d = \frac{d_f}{1 - \sigma \frac{N(d_f - f_0)}{k f_0^2}} \tag{5.59}$$

#### 5.7.2.2 Performance evaluation

To evaluate the performance of the algorithms involved in the depth estimation process, it has been proposed to use the rendering software 3DS Max which enables to create specific stimulus for well-defined conditions: the distance of the camera to the objects, the focal length, the aperture, and the sensor size and resolution. Based on this setting, it was possible to render different images with different amounts of defocus blur. Figure 5.37 depicts an example of a generated stimulus. The blur measurement algorithm and equation 5.59 was then used to recover the depth position defined by design.

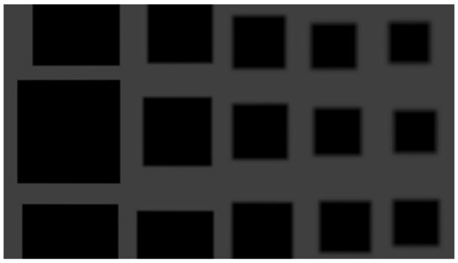


Figure 5.37: Example of stimulus used for evaluating the blur to depth conversion process.

Figure 5.38 depicts the performance of the algorithm. For various values of the aperture the proposed blur measurement from Zhuo [141] can appear noisy and tend to saturate. An alternative measurement algorithm from Chen [162] was considered. Similarly to the Zhuo's method, the proposed method detects the edges using the Canny edge detector. As a second step, for each pixel belonging to the edge the normal to the edge is found and the distribution of the gradient across the edge is studied in the frequency domain. To perform this task, the values of the gradient across the edge are considered as a one-dimensional series which is then analyzed in the frequency domain by applying a fast Fourier transform (FFT). The integral of the power spectrum is then used as a measurement of blur. Figure 5.40 depicts the overall process of blur estimation. The performance of this algorithm appeared to be much more stable and enabled to better estimate the depth values from the defocus blur as it can be seen in Figure 5.39. Therefore it is proposed to use Zhuo's approach to estimate the dense depth map from sparse defocus values, but using the approach from Chen to obtain the sparse blur map.

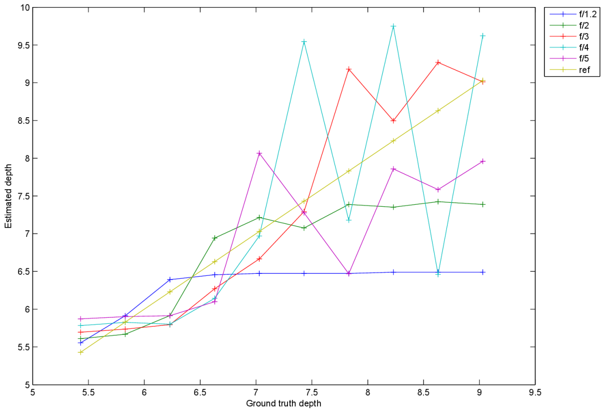


Figure 5.38: Depth from blur with different apertures using the algorithm from Zhuo [141]

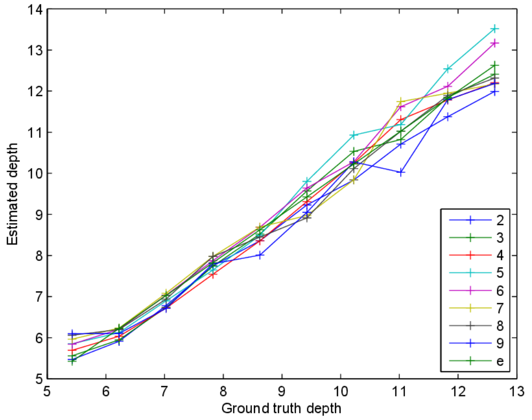


Figure 5.39: Depth from blur with different apertures using the blur measurement algorithm from Chen [162]

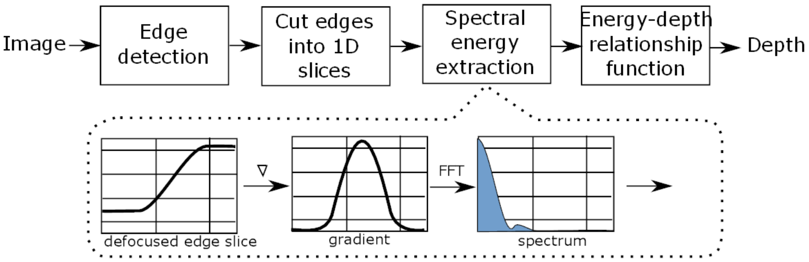


Figure 5.40: Blur estimation process from Chen. Figure copied from [162]

#### 5.7.2.3 Depth cue indicator

In case of natural images, the information about capture settings may not always be available. In many cases, this can be found in the information attached to the picture in the EXIF data. This can be critical in case the depth value is targeted in order to study, for example, the agreement between depth cues. In case of the study of the availability of a depth cue, it can be chosen to only study the variation of the defocus across the dense depth map from defocus blur. As shown in the equation 5.59, the relationship between blur measurements and depth is non-linear therefore it will provide on different results depending on if indicators are built over the blur map or the depth from defocus blur map. Therefore it can only be recommended to document the chosen approach. To evaluate the performance of the depth of defocus blur indicators, the predictions of the algorithm are compared to the subjective ratings obtained in Section 4.5.3. Figure 5.41, depicts the relationship between the minimum amount of blur and the defocus depth cue. As expected, it can be seen that increasing the overall sharpness of the picture decreases the contribution of the depth from defocus: if there are no blurred areas, then there is no depth cue from defocus blur. However, having blurred areas does not necessarily result in having a strong depth from defocus depth cue: the picture can just appear to have blurred areas. To develop a depth cue indicator, different factors have been considered: the percentiles 1% and 90% in order to consider the range of sharpness in the picture. The proportion of sharp areas compared to the blurred area is also measured by using the sparse depth maps and the number of pixels on which it was possible to find edges and compute a blur values. From the data, it was observed that the square of the percentiles 1% and 90% should also be considered. A linear regression between these three factors was performed, and the performance of the proposed indicator is depicted in Figure 5.42. As already found, the effect of minimum blur amount enables to define a lower bound for the prediction of depth from defocus. The challenge still remaining is to detect if the presence of blur contribute to the depth of defocus. Looking into the range of the amount of blur, and then presence of sharp and blurred areas is a way to dig into this question. However, as depicted in Figure 5.42, it was not enough to fully address the problem.

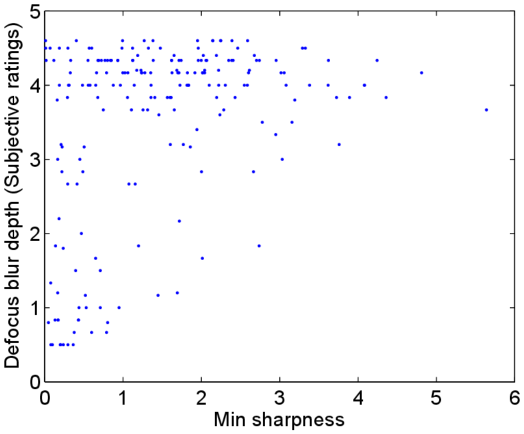


Figure 5.41: Relationship between minimum sharpness and subjective rating for defocus blur depth cue.

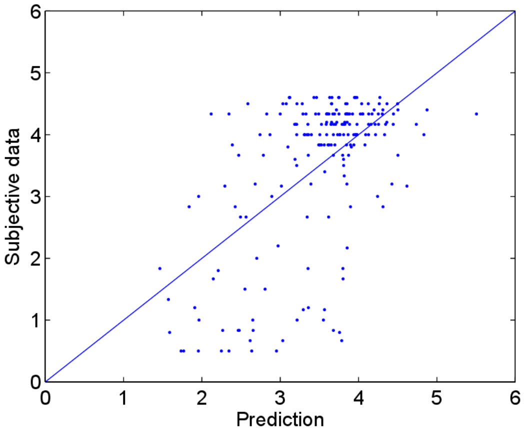


Figure 5.42: Performance of proposed depth indicator against subjective scores from Section 4.5.3.

### 5.7.3 *Motion parallax*

The motion parallax is the difference in apparent motion between objects at different distances in depth. Figure 5.44 depicts two examples, of motion parallax. On the left side, it is possible to distinguish a decrease in the amount of motion in function of the distance to the viewer. On the right side, no such thing can be found, hence no motion parallax contribute to the understanding of the depth. To capture this, a new metric was designed to evaluate the amount of motion parallax. As depicted in Figure 5.44 motion parallax is a difference in apparent motion between object at different distances in depth. It is then suggested to use binocular disparities to estimate the position in depth of the objects and then relate this position in depth to the amount of motion which could be observed (Figure 5.43 depicts examples of dense depth map and dense optical flow revealing the presence of motion parallax). If there is motion parallax, the motion should decrease in function of the depth. Binocular disparities are estimated as explained in Section 5.3.1, a dense optical flow is also estimated similarly. Once the depth map and optical flow are available, the second step of the algorithm is to discard the parts of the images where the maximum value of disparities is reached. This is motivated by a concern of robustness: the algorithm target to relate planar motion and position in depth, but considering that disparity values are used to estimate the depth, object too far away will have a constant depth value and these areas may or may not show motion parallax. It is then proposed to limit the study to areas where motion and depth are clearly known. To limit noise, for each depth plane, the average value of motion of pixels within the depth plane is determined. Figure 5.45 depicts three squatter plots which represent how the average motion change with the disparity values. The two images corresponding to the scatter plots on the right of the figure are images which shows motion parallax: a clear link between the average motion and the position in depth can be found. The images corresponding to the left-bottom part of the figure do not show variation of motion with the variation of depth. The images corresponding to the left-top does not show a linear relation between binocular disparity and motion, this case is similar to Figure 5.44 left where no motion parallax is visible. The RANSAC algorithm is then used with a linear model to fit the relationship between binocular disparities and motion. After fitting, the case of the left-top scatter plot of Figure 5.45 can be distinguished from the other cases due to the low performance of the fitting. In this specific case, the algorithm decides that no motion parallax can be found in the image. The motion parallax is then defined as $MP = \alpha = atan(\frac{dy}{dx})$ with $\frac{dy}{dx}$ the slope of the fitting curve determined by using RANSAC. This $\alpha$ value corresponds to the $\alpha$ defined in Figure 5.44.

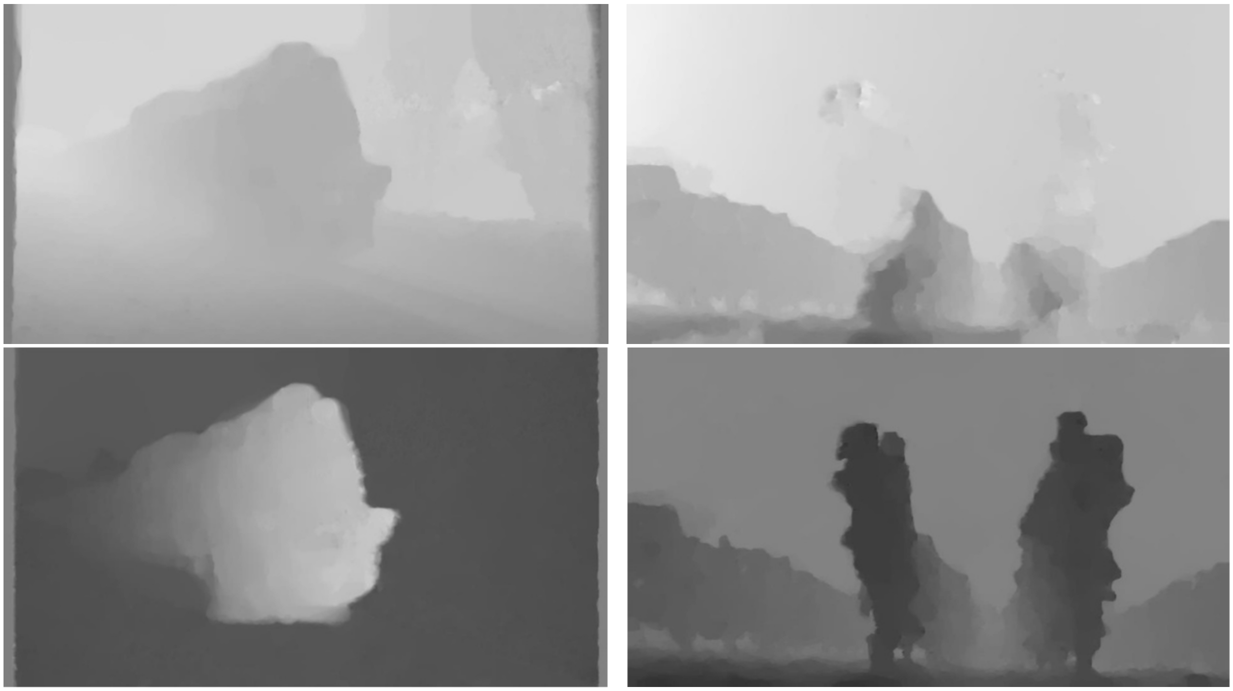


Figure 5.43: Evaluating motion parallax based on dense depth map, and dense optical flow. The top images depicts estimated dense depth map, and the bottom ones dense optical flow.

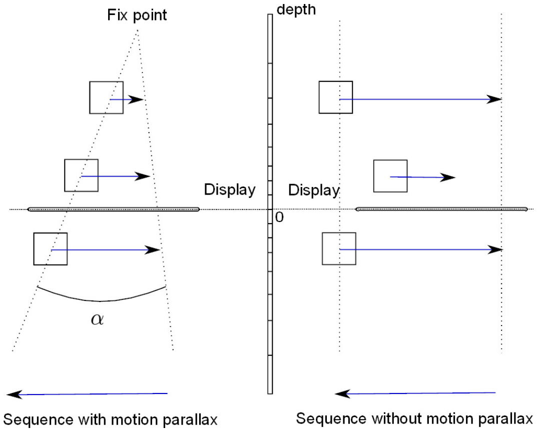

Fix point
depth
Display
Display
0
α
Sequence with motion parallax
Sequence without motion parallax

Figure 5.44: Illustration of motion parallax

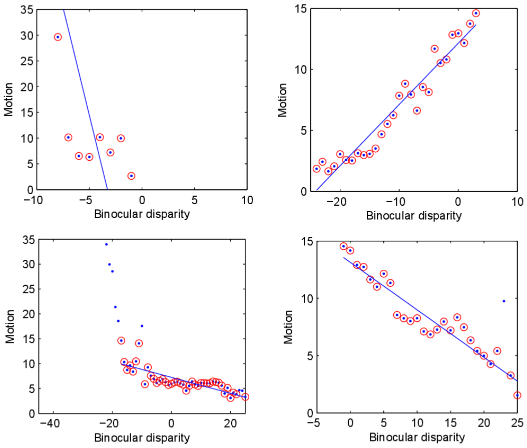


Figure 5.45: Scatter plot motion as a function of disparity on four different images. Fitting and outliers removal are obtained by employing RANSAC, inliers are circled in red

### 5.7.4 Texture gradient

The texture gradient, as described in Section 2.2.2.2, can provide information on the perceived depth due to different factors. These are divided into three categories: *perspective*, *compression* and *density* [39]. *Perspective*: due to the linear perspective, the size of the objects decreases with the distance to the observer, therefore the size of the individual texture element (texel) is affected by this phenomenon. *Compression*: relates to the ratio between the width and height of the texels. The aspect ratio of the texels will be affected by the position. Finally, *density* refers to the spatial distribution of the texels in the image.

Estimating depth from these different factors is a difficult challenge which goes beyond the work of this thesis. Therefore, the algorithm from Agrawal et al [143] was used to recover shape from texture, and provides a depth map from the texture analysis. Once this depth map obtained, a similar challenge as described in Section 5.7.2.3 is raised: the establishment of a single indicator summarizing a dense depth map. It was then decided to employ a consistent analysis of the depth map to the one performed to depth from defocus. Therefore the percentiles 1% and 90% of the depth maps values were considered as well as the standard deviation. A multiple regression between these factors and the subjective data from Section 4.5.3 is then used to find weighting between the percentiles and the standard deviation of depth values measured using the shape from texture algorithm. This linear combination will then be used in the further steps of the thesis.

## 5.8 Depth cues pooling and reliability

To model global depth perception from individual depth cues, it is necessary to consider two main aspects. The first one is that each individual depth cues do not contribute equally to the global depth score, the second and crucial aspect is the confidence in each metric. Indeed all metrics have a specific scope of application and might not always estimate the features correctly. It is then needed to define a weighting factor for each metric. This factor is based on the estimation of the reliability of each individual metric considering the specific image under study. Different approaches can be considered to address this issue and will be described in this section.

### 5.8.1 Reliability and temporal consistency

In case a week fusion model is considered (see chapter 2, section 2.3.2), the depth cues can be combined linearly. Therefore the integration of depth cue reliability can be performed as described in equation 5.60. With $GD_t$ the global depth score of a frame $t$, $ND$ the number of depth cues, $DC_{k,t}$ the depth cue score for the frame $t$ by the metric described in sections 5.4 and 5.7, $cw_k$ a confidence weighting factor, and $pw_k$ a weighting of the contribution of the depth cue $k$ compared to the others.

$$GD_t = \sum_{k=1}^{ND} cw_k \times pw_k \times Dc_{k,t} \tag{5.60}$$

The question of the reliability of the different metrics was addressed in [61]. Different approaches were considered to perform the pooling of two depth cues. The most effective one was found to be the *maximum likelihood estimation model* (MLE). This approach considers the variability of the subjective scores to define a weighting of each depth cue (eq. 5.61).

$$cw_k = \frac{\sum_{i=1,i\neq k}^{ND} \sigma_{DC_i}^2}{\sum_{i=1}^{ND} \sigma_{DC_i}^2} \tag{5.61}$$

A direct application of this approach can be applied to this study by considering the temporal variation of the evaluated depth cues in video sequences. Indeed, it is expected to have, at least locally, a temporal consistency of the evaluated depth scores. Others approaches based on in-depth analysis of each metric are also under study and should be considered since the depth score values can be stable but incorrect. However this will not be addressed in this section. Here only the temporal consistency of each depth cue on a temporal window will be checked. This temporal window is also designed to be constrained to a scene. The confidence metric become as expressed in eq. 5.62. With $cw_{k,w}(t)$ the confidence of the depth cue $k$ for the frame $t$ considering a window $w$.

$$cw_{k,w}(t) = \frac{\sum_{i=1,i\neq k}^{ND} \sigma^2_{DC_{i,w}(t)}}{\sum_{i=1}^{ND} \sigma^2_{DC_{i,w}(t)}} \tag{5.62}$$

#### 5.8.1.1 Ground truth database

As an example, such approach was applied on the database previously defined in section 4.4 which consists of a database composed of 64 10s long video sequences which were displayed at the highest quality available and contained no visible compression artefact. Three scales were asked to the observers: the depth quality, the QoE and the visual discomfort.

Figure 5.46 depicts the results of the evaluation of the different depth cue on one particular video sequence. This sequence is composed of two distinct scenes. Both scenes are natural and shot outdoor. The first one shows a close-up of a hand collecting grapes on a grapevine. The scene is static and only the leaves of the grapevine are moving due to the wind and the hand cutting out the grapes. The second scene shows a lot of grapevines in line until the horizon. The camera moves laterally producing a pan. A clear motion parallax is then visible with the line of trees. The lines of trees allow seeing a clear vanish point in the center of the image. On Figure 5.46 it is possible to perceive a case of failure in one of the metric: the VPM did not always succeed to extract the vanishing lines and has resulted in high temporal variation of this depth cue. This metric is not trustworthy in this specific sequence and should be then discarded. Interestingly it is also possible to see that the second approach for evaluating linear perspective, the GLP, succeeds to evaluate the linear perspective. This metric based on a frequency analysis of the image is better suited to these specific cases of images where no clear edges can be found. A second case of failure can be observed with the motion parallax metric: in the first scene, lots of variation is visible. However in the second scene, the motion parallax is successfully captured and the evaluation is more stable. The proposed approach based on the MLE is then able to capture explicit case of failure in the different metrics.

### *5.8.2 Identification of cases of failure*

The temporal consistency is only one criteria to identify the case of failure, other criteria needs to be taken into account enabling to detect cases where temporal stability can be reached but with a wrong prediction. The particular case of the linear perspective depth cue will be presented in this section.

#### 5.8.2.1 Type of image content

As described in Section 5.7.1, the performance of both algorithms VPM and GLP have been found to have lower performance in case of images classified as “natural”, than in the case of “urban” scenes. Therefore the identification of types of image content can provide a first indication on whether the characterization is likely to be reliable. Using the GIST features from Oliva [159] and a cluster weighted model as suggested in [158], it is proposed to identify the types of scenes: “natural”, or “urban” (Figure 5.47). The performance of the scene-type classification is depicted in Figure 5.48 in a precision/recall diagram. The cluster weighted model provides continuous scores from 0 to 1, 0 for an urban scene and 1 for a natural scene. The Figure 5.48 depicts then the precision and recall for different values of the threshold making the separation between natural and urban scenes. The performance of the scene classifiers is good

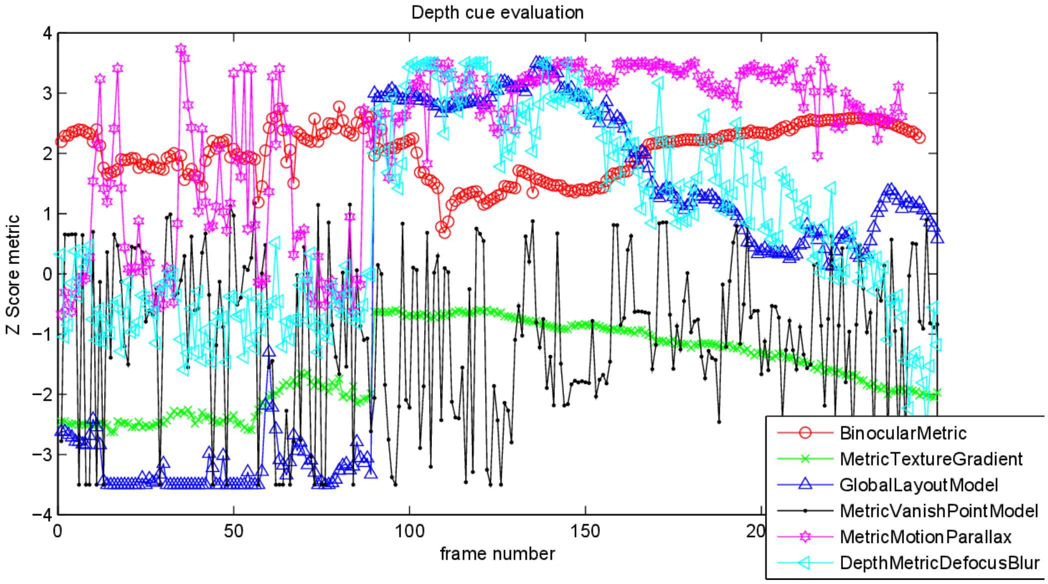


Figure 5.46: Example of depth cues evaluation on a video sequence. The z-scores are determined based on the mean and standard deviation of each score over the entire database

and enable to detect the type of scene content. This can further be used to identify the cases of natural images where the VPM can appear too unstable.

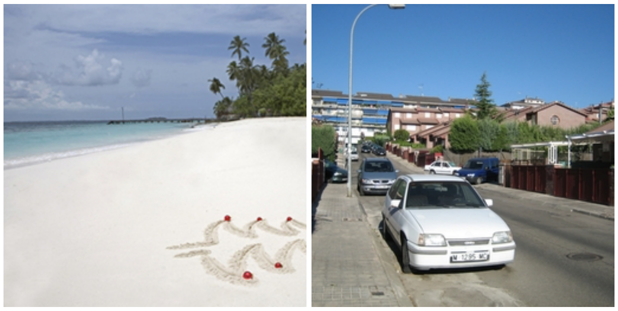


Figure 5.47: Urban or Natural ?

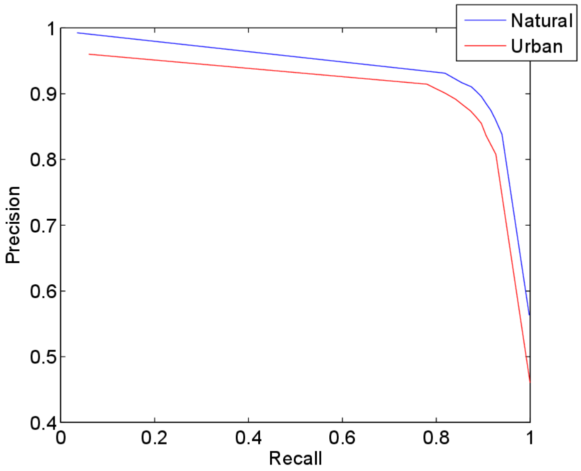


Figure 5.48: Performance of scene type classification using the set of GIST features.

#### 5.8.2.2 Number of vanishing points found

In the process of estimating the vanishing point in the images, the algorithm LSD [161] was used. It enables to get different strokes which are then classified by the J-Linkage algorithm. Figure 5.49 depicts different cases of extracted strokes. Different cases are visible, on the left picture only small strokes have been detected resulting in less stable vanishing point estimation. On the contrary, on the last right picture the extracted strokes are long resulting in a less noisy vanish point estimation. Figure 5.50 depicts an example of small strokes resulting in noisy vanishing point. The figure shows the location of the vanishing point at different resolutions. In this particular example, the vanishing point which will be considered by the algorithm is consistent with what would have been expected, but a non-parametric KruskalWallis one-way analysis of variance shows that the factor “number of vanishing points” affects significantly the prediction error of the algorithm (p<0.01, Chi-sq=179) (see Figure 5.51).

#### 5.8.2.3 Sum on reliability evaluation for linear perspective instrumental measurements

In this subsection the question of identifying cases of failure was considered. Two different approaches were suggested: by extracting features about the image, by the means of content classification and recognizing the kind of picture, natural or urban, which are more error-prone than the other. The second approach looked into the intermediate steps of the prediction algorithms: it relates the length of extracted strokes and number of intermediate vanishing points to the performance of metrics. In both cases, this provides an indication of the accuracy of the prediction algorithm which can further be used as a weighting in equation 5.61, or an indication of the confidence in the content characterization.

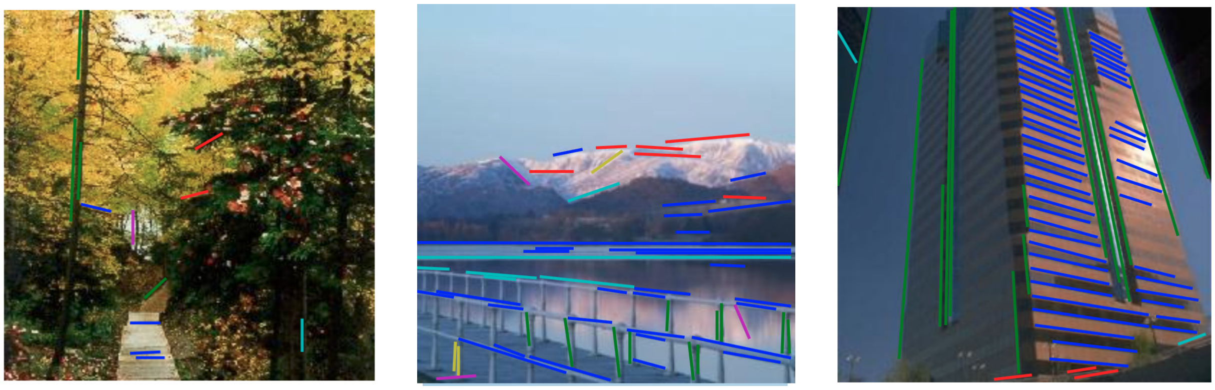


Figure 5.49: Detection of vanishing lines.

### *5.8.3 Outcomes on reliability measurements*

In this section, the issue of depth cue pooling was addressed. Similarly as the work conducted using subjective evaluation methods, it is proposed to also take into account the reliability of instrumental measurements during pooling or more generally when using a depth cue indicator based on algorithms. Several approaches were considered to perform such a task, first temporal consistency of the metric was expected and can contribute to identify cases of failure in depth cue prediction. Secondly, it was proposed on one specific example: the linear perspective metric. the study of its case of failure was based on two distinct levels:

- The recognition of the image properties: natural vs. urban

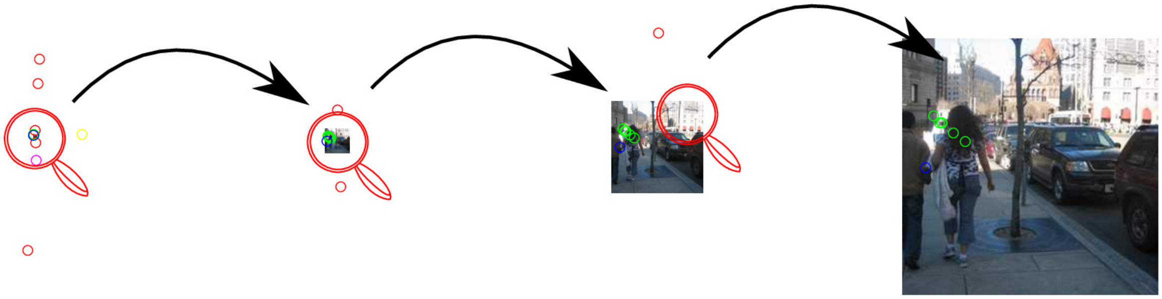


Figure 5.50: Example of multiple vanishing point found due to several small strokes size

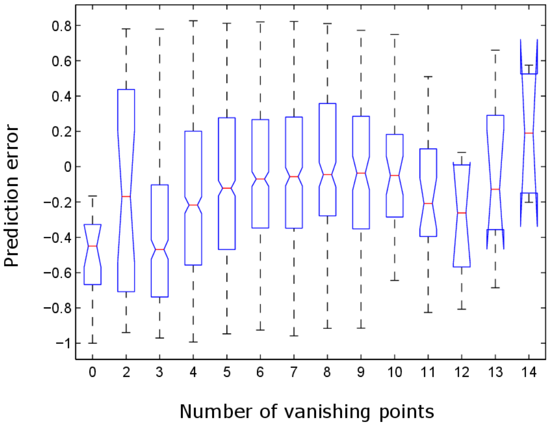


Figure 5.51: Distribution of VPM prediction error depending on the number of vanishing point extracted into the image

- The study of parameters extracted from intermediate steps of the algorithm, e.g. the number of vanishing point and length of vanishing lines.

This enables to better identify if, in the context of use, the prediction algorithm will be reliable or not, and improve the process of decision-making based on these indicators.

## 5.9 Conclusion on depth characterization

In this chapter the question of characterizing the properties of 3D video sequences was studied. Both monocular and binocular depth cues were considered, and depth indicators were developed to evaluate different monocular cues:

- The binocular depth cues
- The linear perspective.
- The defocus blur
- The texture gradient

- The motion parallax

Only little research has been done on the topic of characterizing monocular and binocular depth cues in natural images for 3D video sequences. Therefore it was only possible to compare the proposed algorithm to a limited number of other prediction algorithms. Even if there is a lot of space for improvement, these algorithms provided a novel way to characterize 3D video sequences.

Considering that the performance of the proposed algorithms can vary depending on the type of image content, it was proposed to study the performance of the metric to determine the confidence in the prediction of the algorithm. This was done using either an analysis of temporal consistency of the metrics, or by studying the intrinsic properties of the images or finally by studying parameters on intermediate steps of the metrics enabling to have a measure of the prediction accuracy.

These metrics can be used for different purposes, 3D content classification, QoE models, 3D content selection for subjective testings, depth modeling, etc. In this chapter, a first analysis of depth perception as a function of monocular and binocular depth cues in natural images was also provided. The relation found is: monocular properties of the images affected the binocular depth scores. However, considering the relation between the monocular scores and the binocular measurements though instrumental measurements, the conclusion is rather that the relative size and linear perspective depth cues affect the intrinsic properties of the images and then implies multiple depth layers which induce a higher perceived depth.

Finally, considering the performance of the different monocular and binocular characterization, it has to be mentioned that it is not yet possible to train a weak model for depth perception based on the depth cue evaluation metrics.

Nevertheless, considering the current lack of standardized algorithms for 3D content characterization, the result of this thesis and the different metrics developed can be used for further analysis on content characterization and classification. The code of all the different metrics developed along this thesis were published and freely available for further research [163].

## 5.10 Key contributions

The key contributions of this chapter are the following:

- Monocular characteristics of the images affected the binocular depth scores. However it is not clear whether the monocular depth cues themselves affected the depth percept or if the presence of these depth cues implies specific image properties which change the depth percept
- Binocular and Monocular depth cue indicators were developed to evaluate different cues: binocular depth cues, linear perspective, defocus blur, texture gradient, motion parallax
- The question of depth indicator reliability was addressed across different methods of measurements: using temporal consistency, and features extracted on the studied images and on the metrics while estimating the cue value.
- Last but not least, all the code of each individual metric have been published enabling further studies to reuse the work performed along this thesis for content characterization. The metric can be accessed at the following doi: http://dx.doi.org/10.5281/zenodo.16925

# Chapter 6
# Conclusion

The work performed in this thesis addresses different issues. It starts from the evaluation of Quality of Experience in 3D video sequences, providing an overview of the related literature on QoE and visual depth perception. It was observed that when test participants are asked to rate Quality of Experience, their ratings depend on their test-specific concept of QoE. In particular, it was observed that they do not necessarily provide consistent ratings across test participants, and use the scales (QoE, Visual comfort, and Depth) differently. The issue of consistency, agreement between test participants, and the understanding of the scales by the test participants has been studied along this thesis.

The first test results obtained in this work showed, that test participants do not necessarily rate 3D to provide a higher QoE than 2D, and therefore do not necessarily take into account the added value that 3D may provide in their ratings: The availability of binocular depth cues. To overcome this problem, the paired comparison test paradigm has been used to evaluate the preference of different 2D and 3D stimuli, extending respective work on this topic presented in the literature. Using this method, it has been possible to show the preference of 3D over 2D in specific conditions, and to quantify the added value of 3D.

The added value of 3D was found to be content-dependent. The preference of 3D over 2D was found to decrease with a decrease of image quality. This decrease of preference depends on the content properties. To evaluate 3D-QoE, there thus is a need to *characterize* 3D video sequences. Since the added value of 3D is to bring binocular depth, the work has been focused on the evaluation of depth, starting with depth in natural images. First, depth was assessed in viewing tests, in a second step using prediction algorithms. Considering the fact that depth perception results from different monocular and binocular depth cues, different tests involving test participants have been conducted to evaluate the depth in images and videos. However, the question of subjective scores' reliability and agreement between test participants was raised. It was shown that test participants do not necessarily understand the different depth scales in the same manner. Therefore the research effort of this thesis was focused on defining and assessing depth cues. A series of studies has been conducted to evaluate subjective test methods and develop a simple way for test participants to evaluate monocular and binocular depth cues in natural images. Different approaches using pairwise comparison and ranking of images have been proposed were shown to enable the acquisition of rating data with increased reliability.

Based on the subjective scores obtained in these tests, new prediction algorithms were designed to characterize the properties of 3D video sequences: the overall perceived depth, and different underlying monocular and binocular depth cues. An algorithm to predict the perceived depth from binocular depth cues performing an object-based analysis of the scene was established enabling to monitor 3D content properties in videos. Additionally different monocular depth cues indicators for defocus blur, linear perspective, texture gradient and motion parallax were defined. The accuracy of the prediction algorithms was found to not always be optimal. Therefore, similarly to the analysis of the data collected from test participants, it has been proposed to study the performance and trust in the different metrics. It has been proposed to study different aspects such as the temporal consistency, image classification, and features on the metrics to enable quantifying the prediction accuracy.

The development of these metrics was of particular interest since until now no standardized way to characterize the properties of 3D contents have been proposed. One major contribution of this thesis is therefore the open-source publication of all of these indicators enabling further research to characterize the content properties with the algorithms developed in this thesis.

From a visual perception point of view, it was difficult to draw strong conclusions about the depth perception and the relation between monocular and binocular depth cues. A relationship between monocular and overall depth perception could be found. However, from these data it is not possible to conclude to which extent the monocular depth cues contribute to the overall depth perception or if these depth cues affect the image-intrinsic properties which then affect the overall perception of the scene.

# Chapter 7
# Further work

Across the thesis, different aspects have been addressed from visual perception studies to content characterization including Quality of Experience research, and depth perception prediction. The final goal of the thesis is to address Quality of Experience and its relation to content properties, that is, the contents' depth properties. As a consequence, each of the addressed aspects from the perception studies to image analysis were fundamental. Due to the variety of topics which have been addressed, there are several possibilities for further research along different axis.

***Quality of Experience research:*** On the topic of Quality of Experience research, further work could consider to relate content characterization performed in this thesis and the preference of 3D over 2D. The work described in this thesis have related content, image quality (due to coding) and preference of 3D over 2D. We have also found that 2D image quality algorithms performed relatively well to predict 3D image quality (with traditional 2D coding schemes). Moreover, algorithms for the content characterization were also proposed. However, the link between the subjective ratings from our experiments on preference of 3D over 2D and the content properties and respective quality requirements as a function of the proposed content-specific depth descriptor still needs to be continued, based on an even larger set of ratings. Analysis have been performed in this direction within the thesis, but this requires a large amount of subjective test as many source content are needed. In this work, it was then decided to focus on content characterization.

***Content characterization research:*** Another extension of this work, could be to take into account the different depth descriptors defined in this work, analyze the agreement and contradiction between them, and in addition measure distortions of the 3D geometry of 3D contents. Some work has been performed in this direction in the literature for the case that camera parameters are known (Chen [12]), however our approach through image processing would provides a new way to look at this issue and predict QoE of source content without prior knowledge of camera settings. To this aim, instead of 3D QoE assessment in terms of preference, the different depth cue indicators produced in this work could be evaluated in terms of how well they allow to measure the depth quality of the 3D rendering. For example, this could be used to evaluate the cardboard effect. It could also be interesting to study when depth cues contradict with other. This could be used for further analysis of depth cue prediction reliability, but it also could be used to address visual discomfort issues.
The characterization of monocular depth cues can also be used beyond 3D. Indeed, it could be used to measure aesthetic appeal in pictures. For example, by considering vanishing lines, position of the vanishing points and common rules on photography. Defocus blur metrics could also be used in the context of Ultra High Definition contents: evaluating the distribution of the sharpness across the images, and relate it to what observers may perceive in the side areas of their field of view.
Finally, other depth cues not instrumentally characterized in this thesis need to be addressed in future research.

***Perception studies:*** Last but not least, the perception studies on natural images was one of the most challenging topic addressed in this thesis, and could be extended. To enable stronger conclusions, it is proposed to extend the work to content designed using 3D rendering software. The use of natural images in this thesis has provided some insights on the depth perception and the relative importance of each depth cue. However, it would be beneficial to also consider content where only one depth cue is variated at a time instead of evaluating natural images as done in this thesis, where different depth cues are present at the same time requiring to also evaluate each depth cue in addition to the overall depth. This will enable to dig deeper into the construction of a 3D vision model based on monocular and binocular depth cues as it has been stated in this thesis.

# References

1. Patrick Le Callet, Sebastian Möller, and Andrew Perkis, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, vol. Version 1.2, March 2012.
2. Pieter J.H. Seuntiëns, *Visual experience of 3D TV*, Ph.D. thesis, Eindhoven University, 2006.
3. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "A subjective evaluation of 3D IPTV broadcasting implementations considering coding and transmission degradation," in *IEEE International Workshop on Multimedia Quality of Experience, MQoE11*, Dana Point, CA, USA, 2011.
4. Kun Wang, Marcus Barkowsky, Kjell Brunnström, Marten Sjöström, Romain Cousseau, and Patrick Le Callet, "Perceived 3D TV transmission quality assessment: Multi-laboratory results using Absolute Category Rating on Quality of Experience scale," *IEEE Transactions on Broadcasting*, vol. 58, pp. 544–557, 2012.
5. Wijnand IJsselsteijn, Huib De Ridder, Jonathan Freeman, and S. E. Avons, "Presence: Concept, determinants and measurement," in *Proceedings of the SPIE*, 2000, vol. 3959, p. 520529.
6. Wijnand Ijsselsteijn, Pieter J.H. Seuntiëns, and L. Meesters, "ATTEST Deliverable1: State of the art in human factors and quality issues of stereoscopic broadcast television," Tech. Rep., Eindhoven University of Technology, 2002.
7. Wijnand Ijsselsteijn, Huib De Ridder, Jonathan Freeman, S. E. Avons, and Don Bouwhuis, "Effects of Stereoscopic Presentation, Image Motion, and Screen Size on Subjective and Objective Corroborative Measures of Presence," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 298–311, 2001.
8. Pierre Lebreton, Marcus Barkowsky, Alexander Raake, and Patrick Le Callet, *Chapter 3D Video, "Quality of Experience - Advanced Concepts, Applications and Methods"*, Springer, 2014.
9. Marc Lambooij, Wijnand IJsselsteijn, Don G. Bouwhuis, and Ingrid Heynderickx, "Evaluation of Stereoscopic Images: Beyond 2D Quality," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 432444, 2011.
10. Lew Stelmach, Wa James Tam, Dan Meegan, and André Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 2, pp. 188 –193, march 2000.
11. Kazuhisa Yamagishi, Lina Karam, Jun Okamoto, and Takanori Hayashi, "Subjective characteristics for stereoscopic high definition video," in *Third International Workshop on Quality of Multimedia Experience, QoMEX*, Mechelen, Belgium, 2011.
12. Wei Chen, Jérôme Fournier, Marcus Barkowsky, and Patrick Le Callet, "Exploration of quality of experience of stereoscopic images: binocular depth," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, USA, 2012.
13. Atanas Boev, Danilo Hollosi, Atanas Gotchev, and Karen Egiazarian, "Classification and simulation of stereoscopic artifacts in mobile 3DTV content," in *Proc. SPIE 7237, Stereoscopic Displays and Applications XX*, 2009, vol. 7237.
14. Randolph Blake, "Threshold conditions for binocular rivalry," *Journal Experimental Psychology Human Perception Performance*, vol. 3, no. 2, pp. 251–257, 1977.
15. Marc Lambooij, Wijnand IJsselsteijn, Marten Fortuin, and Ingrid Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53(3), pp. 1–14, 2009.
16. Jong-Seok Lee, Lutz Goldmann, and Touradj Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia Tools and Applications*, pp. 1–18, February 2012.
17. *A comprehensive Database and subjective evaluation methodology for quality of experience in stereoscopic video*, 2010.
18. Peter G. Engeldrum, "Image quality modeling: Where are we?," in *IS&T PICS Conference Proceedings*, 1999, pp. 251–255.
19. Marc Lambooij, *Visual Comfort of 3-D TV MModel and Measurements*, Ph.D. thesis, Eindhoven University of Technology. Departement of Industrial Engineering and Innovation Sciences. Human-Technology Interaction Group, 2012.
20. Heinrich H. Bülthoff and Hanspeter A. Mallot, "Integration of depth modules: stereo and shading," *Journal of the Optical Society of America*, vol. 5, no. 10, pp. 1749–1758, October 1988.
21. Hirokazu Yamanoue, Makoto Okui, and Ichiro Yuyama, "A study on the relationship between shooting conditions and cardboard effect of stereoscopic images," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 3, pp. 411 – 416, 2000.
22. Hirokazu Yamanoue, Makoto Okui, and Fumio Okano, "Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images," *IEEE Transaction on circuits and systems for video technology*, vol. 16, pp. 744 – 752, 2006.
23. Jae-Hyun Jung, Jiwoon Yeom, Jisoo Hong, Keehoon Hong, Sung-Wook Min, and Byoungho Lee, "Effect of fundamental depth resolution and cardboard effect to perceived depth resolution on multi-view display," *Optics Express*, vol. 19, no. 21, pp. 20468–20482, 2011.
24. Wei Chen, Jérôme Fournier, Marcus Barkowsky, and Patrick Le Callet, "New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone," *Stereoscopic Displays and Applications XXII. Proceedings of the SPIE*, vol. 7863, pp. 78631O–78631O–13, 2011.
25. James E. Cutting and Peter M. Vishton, *Perceiving layout and knowing distance: The integration, relative potency and contextual use of different information about depth*, New York: Academic Press, 1995.
26. Neil A. Dodgson, "Variation and extrema of human interpupillary distance," in *Proceedings SPIE Vol. 5291, Stereoscopic Displays and Virtual Reality Systems XI*, 2004, p. 3646.
27. R. Patterson and W. L. Martin, "Human stereopsis," *Hum. Factors*, vol. 34, pp. 669692, 1992.
28. P. Howard and B. J. Rogers, *Seeing in Depth: Depth Perception*, vol. 2, Porteous Publishing, Toronto, 2002.
29. Y. Y. Yeh and L. D. Silverstein, "Limits of fusion and depth judgement in stereoscopic color displays," *Hum. Factors*, vol. 32, pp. 4560, 1990.

30. C. Schor, I. Wood, and J. Ogawa, “Binocular sensory fusion is limited by spatial resolution,” *Vision Research*, vol. 24, pp. 661665, 1984.
31. Steven H. Ferris, “Motion parallax and absolute distance,” *Journal of experimental psychology*, vol. 95, no. 2, pp. 258–263, 1972.
32. “http://en.wikipedia.org/wiki/depth_perception,” .
33. Michael T. Swanston and Walter C.. Gogel, “Perceived size and motion in depth from optical expansion,” *Perception & Psychophysics*, vol. 39, no. 5, pp. 309326, 1986.
34. William H. Ittelson, “Size as a cue to distance: Radial motion,” *American Journal of Psycholoyg*, vol. 64, no. 2, pp. 188–202, 1951.
35. ,” .
36. James E. Cutting, “How the eye measures reality and virtual reality,” *Behavior Research Methods, Instruments, & Computers*, vol. 29, no. 1, pp. 27–36, 1997.
37. Irving B. Weiner, *Handbook of Psychology, Experimental Psychology*, Wiley, 2012.
38. Rita Sousa, Eli Brenner, and Jeroen B. J. Smeets, “Judging an unfamiliar object’s distance from its retinal image size,” *Journal of Vision*, vol. 11, no. 9, pp. 1–6, 2011.
39. Cutting and Millard, “Three gradients and the perception of flat and curved surfaces,” *Journal of Experimental Psychology: General*, vol. 113, pp. 198216, 1984.
40. H Sedgwick, “The visible horizon: A potential source of visual information for the perception of size and distance,” *Dissertation Abstracts International*, vol. 34, no. 73, pp. 13011302, 1973.
41. James Elkins, “Renaissance perspectives,” *Journal of the History of Ideas*, vol. 53, no. 2, pp. 209–230, 1992.
42. Pamela Taylor, *The Notebooks of Leonardo Da Vinci Mass Market Paperback*, The New American Library, 1960.
43. John S. Watson, Martin S. Banks, Claes von Hofsten, and Constance S. Royden, “Gravity as a monocular cue for perception of absolute distance and/or absolute size,” *Perception*, vol. 21, no. 1, pp. 69–76, 1992.
44. Heinrich H. Bülthoff and Hanspeter A. Mallot, “Interaction of different modules in depth perception,” in *Proceedings of 1st International Conference on Computer Vision*, 1987, pp. 295–305.
45. Ian P. Howard and Brian J Rogers, *Binocular Vision and Stereopsis*, Oxford phychology series no. 29. Oxford University Press, 1995.
46. I. P. Howard and W. B. Templeton, *Human Spatial Orientation*, John Wiley and Sons, 1967.
47. Mark Young, Michael S. Landy, and Laurence T. Maloney, “A perturbation analysis of depth perception from combinations of texture and motion cues,” *Vision Research*, vol. 33, no. 18, pp. 2685–96, 1993.
48. James J. Clark and Alan L. Yuille, *Data Fusion for Sensory Information Processing Systems*, vol. 105, The Springer International Series in Engineering and Computer Science, 1990.
49. Michael S. Landy, Laurence T. Maloney, Elizabeth B. Johnston, and Mark Young, “Measurement and modeling of depth cue combination: in defense of weak fusion,” *Vision Research*, vol. 35, no. 3, pp. 389–412, 1995.
50. Ken Nakayama and Shinsuke Shimojo, “Experiencing and perceiving visual surfaces,” *Science*, vol. 257, pp. 1357–1363, 1992.
51. Barbara Anne Dosher, George Sperling, and Stephen A. Wurst, “Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure,” *Vision Research*, vol. 26, no. 6, pp. 973–990, 1986.
52. Nicolas Bruno and James E. Cutting, “Minimodularity and the perception of layout,” *Journal of Experimental Psychology: General*, vol. 17, no. 2, pp. 161–170, 1988.
53. Elizabeth B. Johnston, Bruce G. Cumming, and Andrew J. Parker, “Integration of depth modules: Stereopsis and texture,” *Vision Research*, vol. 33, no. 5/6, pp. 813–826, 1993.
54. David Buckley and John P Frisby, “Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges,” *Vision Research*, vol. 33, no. 7, pp. 919–933, 1993.
55. Michael S. Landy, Laurence T. Maloney, and Mark J. Young, “Psychophysical estimation of the human depth combination rule,” *Sensor fusion III: 3-D Perception and recognition, SPIE Proceedings*, vol. 1383, pp. 247–254, 1991.
56. Brian J Rogers and Thomas S Collett, “The appearance of surfaces specified by motion parallax and binocular disparity,” *Journal of Experimental Psychology Section A: Human Experimental Psychology*, vol. 41, no. 4, pp. 697–717, 1989.
57. Junle Wang, Marcus Barkowsky, Vincent Ricordel, and Patrick Le Callet, “Quantifying how the combination of blur and disparity affects the perceived depth,” in *Proceedings of the SPIE. Human Vision and Electronic Imaging XVI*. 2011, vol. 7865, pp. 78650K–78650K–10, Proceedings of the SPIE.
58. Robert T. Held, Emily A. Cooper, and Martin S. Banks, “Blur and disparity are complementary cues to depth,” *Current Biology*, vol. 22, pp. 426–431, 2012.
59. Marc O. Ernst and Martin S. Banks, “Humans integrate visual and haptic information in a statistically optimal fashion,” *Nature*, vol. 415, pp. 429–433, 2002.
60. James M. Hillis, Simon J. Watt, Michael S. Landy, and Martin S. Banks, “Slant from texture and disparity cues: Optimal cue combination,” *Journal of Vision*, vol. 4, pp. 967–992, 2004.
61. Paul G. Lovell, Marina Bloj, and Julie M. Harris, “Optimal integration of shading and binocular disparity for depth perception,” *Journal of Vision*, vol. 12, pp. 1–18, 2012.
62. Dominic W. Massaro, “Ambiguity in perception and experimentation,” *Journal of Experimental Psychology: General*, vol. 117, no. 4, pp. 417–421, 1988.
63. Kenneth N. Ogle, “A new phenomenon in binocular space perception associated with the relative size of the images of the two eyes,” *Archive of Ophthalmology*, vol. 20, no. 4, pp. 604–623, 1938.
64. Myron L. Braunstein, George J. Andersen, and David M. Riefer, “The use of occlusion to resolve ambiguity in parallel projections,” *Perception & Psychophysics*, vol. 31, no. 3, pp. 261–267, 1982.

65. Andrew Blake and Heinrich Bülthoff, "Shape from specularities: computation and psychophysics," *Philosophical Transactions of the Royal Society of London*, vol. 331, no. 1260, pp. 237–52, 1991.
66. Alan L. Yuille and Heinrich H. Bülthoff, "Bayesian decision theory and psychophysics," in *In Perception as Bayesian Inference*. 1994, pp. 123–161, University Press.
67. Ahna R. Girshick and Martin S. Banks, "Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts," *Journal of Vision*, vol. 9, pp. 1–20, 2009.
68. Raymond van Ee, Wendy J. Adams, and Pascal Mamassian, "Bayesian modeling of cue interaction: bistability in stereoscopic slant perception," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1398–1406, 2003.
69. Walter C. Gogel, "An indirect method of measuring perceived distance from familiar size," *Perception & Psychophysics*, vol. 20, no. 6, pp. 419–429, 1976.
70. Elizabeth B. Johnston, "Systematic distortion of shape from stereopsis," *Vision Research*, vol. 31, no. 7/8, pp. 1351–1360, 1991.
71. Elaine W. Jin, Brian W. Keelana, Junqing Chen, Jonathan B. Phillips, and Ying Chen, "Softcopy quality ruler method: Implementation and validation," in *Proceeding of SPIE-IS&T Electronic Imaging*, San Jose, CA, USA, 2009, vol. 7242.
72. Kent A. Stevens and Allen Brook, "Integrating stereopsis with monocular interpretations of planar surfaces," *Vision Research*, vol. 28, no. 3, pp. 371–386, 1988.
73. James S. Tittle and Myron L. Braunstein, "Recovery of 3-D shape from binocular disparity and structure from motion," *Perception & Psychophysics*, vol. 54, no. 2, pp. 157–169, 1993.
74. Brian J Rogers and Maureen Graham, "Similarities between motion parallax and stereopsis in human depth perception," *Vision Research*, vol. 22, pp. 261–270, 1981.
75. W. C. Clarke, A. H. Smith, and A. Rabe, "Retinal gradients of outline distortion and binocular disparity as stimuli for slant," *Canadian Journal of Experimental Psychofogy*, vol. 10, pp. 1–8, 1956.
76. Martin S. Banks, Jenny C. A. Read, Robert S. Allison, and Simon J. Watt, "Stereoscopy and the human visual system," *SMPTE Mot. Imag*, vol. 4, no. 121, pp. 24–43, May-June 2012.
77. E. H. Adelson, "Rigid objects that appear highly non-rigid," *Investigate Ophthalmology and Visual Science*, vol. 26, pp. 3–56, 1985.
78. Matthieu Urvoy, Marcus Barkowsky, and Patrick Le Callet, "How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors," *annals of telecommunications*, vol. 68, pp. 641–655, 2013.
79. M. Emoto, T. Niida, and F. Okana, "Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television," *journal of display technology*, vol. 1, pp. 328340, 2005.
80. W. A. IJsselsteijn, P. H. J. Seuntiens, and L. M. J. Meesters, *3D Videocommunication-Algorithms, Concepts and Real-Time Systems in Human-centred Communication*, chapter Human Factors of 3D Display, p. 219234, John Wiley and Sons, 2005.
81. Kazuhiko Uka and Peter A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations," *Displays*, vol. 29, no. 2, pp. 106116, 2008.
82. D. A. Goss and Z. Huifang, "Clinical and laboratory investigations of the relationship of accommodation and convergence function with refractive error. a literature review," *Doc. Ophthalmology*, vol. 86, pp. 349380, 1994.
83. M. Wopking, "Viewing comfort with stereoscopic pictures: an experimental study on the subjective effects of disparity magnitude and depth of focus," *Journal Society for Information Display*, vol. 3, pp. 101–103, 1995.
84. C. Sheard, "The prescription of prisms," *American Journal of Optomety*, vol. 11, no. 10, pp. 364–378, 1934.
85. Yuji Nojiri, Hirokazu Yamanoue, Atsuo Hanazato, and Fumio Okano, "Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic hdtv," in *Proc. SPIE 5006, Stereoscopic Displays and Virtual Reality Systems X 195*, May 2003.
86. Sumio Yano, Masaki Emoto, and Tetsuo Mitsuhashi, "Two factors in visual fatigue caused by stereoscopic HDTV images," *Displays*, vol. 25, pp. 141–150, 2004.
87. Filippo Speranza, Wa James Tam, Ron Renaud, and Namho Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," in *Stereoscopic Displays and Virtual Reality Systems XIII*, 2006, vol. 6055.
88. ITU-R Recommendation BT.1438, "Subjective assessment of stereoscopic television pictures," 2000.
89. Wei Chen, Jérôme Fournier, Marcus Barkowsky, and Patrick Le Callet, "New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone," *Stereoscopic Displays and Applications XXII. Proceedings of the SPIE*, vol. 7863, pp. 78631O–78631O–13, 2011.
90. Bernard Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema From Scrip to Screen*, Focal Press, 2009.
91. Andrew Woods, Tom Docherty, and Rolf Koch, "Image distortions in stereoscopic video systems," in *Proceedings of the SPIE, Stereoscopic Displays and Applications IV*, 1993, vol. 1915, pp. 36–48.
92. Wei Chen, Jérôme Fournier, Marcus Barkowsky, and Patrick Le Callet, "New requirements of subjective video quality assessment methodologies for 3DTV," in *Video Processing and Quality Metrics 2010 (VPQM)*, Scottsdale, USA, 2010.
93. K.; Merkle P.; Kauff P.; Wiegand T. Smolic, A.; Mueller, "An overview of available and emerging 3d video formats and depth enhanced stereo as efficient generic solution," in *Picture Coding Symposium*, 2009.
94. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "A subjective and objective evaluation of a realistic 3D IPTV transmission chain," in *Packet Video Workshop*, Munich, Germany, 2012.
95. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "Perceptual preference of S3D over 2D for HDTV in dependence of video quality and depth," in *IVMSP Workshop: 3D Image/Video Technologies and Applications*, Seoul, Korea, 2013.
96. "http://x264.nl/," .
97. "http://sirannon.atlantis.ugent.be/," .

98. ITU-R BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," 2009.
99. Manuel Werlberger Thomas Pock and Horst Bischof., "Motion Estimation with Non-Local Total Variation Regularization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA*, 2010.
100. Kun Wang, Marcus Barkowsky et al., "Subjective evaluation of HDTV stereoscopic videos in IPTV scenarios using absolute category rating," *EI2011*, 2011.
101. Kjell Brunnström, Inigo Sedano, Kun Wang, Marcus Barkowsky, Maria Kihl, Börje Andrn, Patrick Le Callet, Marten Sjöström, and Andreas Aurelius, "2D no-reference video quality model development and 3D video transmission quality," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, USA, 2012.
102. ITU-R Rec. BT.2021, "Subjective methods for the assessment of stereoscopic 3dtv systems," International Telecommunication Union (ITU), 2012.
103. Marcus Barkowsky, Jing Li, Taehwan Han, Sungwook Youn, Jiheon Ok, Chulhee Lee, Christer Hedberg, Indirajith Vijai Ananth, Kun Wang, Kjell Brunnstrm, and Patrick Le Callet, "Towards standardized 3DTV QoE assessment: Cross-lab study on display technology and viewing environment parameters," in *Stereoscopic Displays and Applications XXIV, Vol: 8648*, San franscisco : United States, 2013.
104. Ulrich Engelke, Yohann Pitrey, and Patrick Le Callet, "Towards an inter-observer analysis framework for multimedia quality assessment," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, Mechelen, 2011, pp. 183 – 188.
105. F. Kozamernik, V. Steinmann, P. Sunna, and E. Wyckens, "SAMVIQ-A New EBU Methodology for Video Quality Evaluations in Multimedia," *SMPTE Mot. Imag.*, pp. 152–160, April 2005.
106. Q Huynh-Thu, MD Brotherton, D Hands, and K Brunnstrm, "Examination of the SAMVIQ subjective assessment methodology," in *Third Inter. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, USA, 2007.
107. "http://3dtv.at/products/player/index_de.aspx," .
108. Alexandre Benoit, Patrick Le Callet, Patrizio Campisi, and Romain Cousseau, "Quality assessment of stereoscopic images," in *IEEE International Conference on Image Processing , ICIP*, San Diego, California, USA, 2008, pp. 1231–1234.
109. ITU-T Contribution COM 12-C192-E, "Comparison of the ACR and PC evaluation methods concerning the effects of video resolution and size on visual subjective ratings," in *ITU*, SG12 Meeting, Geneva, Jan 2011.
110. J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller model of paired comparisons for image quality assessment," in *PICS*, April 2001.
111. Jing Li, Marcus Barkowsky, and Patrick Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in *ICIP*, Orlando, Florida, USA, October 2012.
112. Marcus Barkowsky, Romain Cousseau, and Patrick Le Callet, "Influence of depth rendering on the quality of experience for an autostereoscopic display," in *International Workshop on Quality of Multimedia Experience*, San Diego, California, USA, 07 2009, p. 6.
113. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "Evaluating depth perception of 3D stereoscopic videos," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, pp. 710–720, October 2012.
114. Matthieu Urvoy and et al., "NAMA3DS1-COSPAD1 : Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Fourth International on Quality of Multimedia Experience*, Yarra Valley, July 2012.
115. Roumes C, Plantier J, Menu JP, and Thorpe S, "The effects of spatial frequency on binocular fusion: from elementary to complex images," *Human Factors*, vol. 39, no. 3, pp. 359–373, Sep 1997.
116. ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," 2008.
117. Jr Otto Dykstra, "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics*, vol. 16, no. 2, pp. 176–188, June 1960.
118. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "Measuring perceived depth in natural images and study of its relation with monocular and binocular depth cues," in *Stereoscopic Displays and Applications XXV*, San Francisco, California, USA, 2014.
119. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, "Evaluating complex scales through subjective ranking," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, Singapore, 2014.
120. Liyuan Xing, Junyong You, Touradj Ebrahimi, and Andrew Perkis, "Assessment of stereoscopic crosstalk perception," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 14, no. 2, pp. 326–337, APRIL 2012.
121. Xing Liyuan, You Junyong, Ebrahimi Touradj, and Perkis Andrew, "Factors impacting quality of experience in stereoscopic images," in *Proceedings of SPIE - The International Society for Optical Engineering*, San Francisco, California, USA, 2011, vol. 7863.
122. Lutz Goldmann, Francesca De Simone, and Touradj Ebrahimi, "Impact of acquisition distortions on the quality of stereoscopic images," in *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, USA, 2010.
123. homepage of, "http://mmspg.epfl.ch/3diqa," last access in Januray 2014.
124. Karel Fliegel, Stanislav Vítek, Milos Klíma, and Petr Páta, "Open source database of images DEIMOS: high dynamic range and stereoscopic content," in *Proc. SPIE 8135, Applications of Digital Image Processing XXXIV, 81351T*, September 2011.
125. homepage of, "http://www.deimos-project.cz/tag/stereo," last access in Januray 2014.
126. E. Cheng, P. Burton, J. Burton, A. Joseski, and I. Burnett, "RMIT3DV: Pre-announcement of a creative commons uncompressed hd 3D video database," in *Proc. 4th International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, Yarra Valley, Australia, 2012.
127. homepage of, "http://www.rmit3dv.com/download.php," last access in Januray 2014.

128. homepage of, "http://www.elephantsdream.org/," last access in Januray 2014.

129. J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973.

130. Berthold Klaus Paul Horn and Brian G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

131. Manuel Werlberger, *Convex Approaches for High Performance Video Processing*, Ph.D. thesis, Graz University of Technology, Institute for Computer Graphics and Vision, 2012.

132. C Schnörr, "Segmentation of visual motion by minimizing convex non-quadratic functionals," *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, pp. 661663, 1994.

133. J. Weickert and C. A. Schnrr, "Theoretical framework for convex regularizer in pde-based computation of image motion," *International Journal of Computer Vision*, vol. 45, no. 3, pp. 245–264, 2001.

134. L. Alvarez, J. Esclarn, M. Lefebure, and J. Snchez, "A pde model for computing the optical flow," *Proceedings XVI Congreso de Ecuaciones Diferenciales y Aplicaciones C.E.D.Y.A. XVI*, vol. 1, pp. 13491356, 1999.

135. I. Gheta, C. Frese, M. Heizmann, and J. Beyerer, "A new approach for estimating depth by fusing stereo and defocus information," in *INFORMATIK 2007: Informatik trifft Logistik. Band 1. Beitrge der 37. Jahrestagung der Gesellschaft fr Informatik e.V. (GI)*, 2007, pp. 26–31.

136. M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 Optical Fow," in *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, September 2009.

137. "http://vision.middlebury.edu/stereo/eval/," .

138. Pina Marziliano, Frederic Dufaux, Stefan Winkler, and TouradjEbrahimi, "A no-reference perceptual blur metric," in *International Conference on Image Processing*, 2002.

139. Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," in *SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*, San Jose, 2007.

140. Jorge Caviedes and Sabri Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *International Conference on Image Processing (ICIP)*, 2002, vol. 3, p. 5356.

141. Shaojie Zhuo and Terence Sim, "Defocus map estimation from a single image.," *Pattern Recognition*, vol. 44, no. 9, pp. 1852–1858, 2011.

142. Anat Levin, Dani Lischinski, and Yair Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 30, no. 2, pp. 228–242, 2008.

143. A. Agrawal, R. Chellappa, and R. Raskar, "An algebraic approach to surface reconstructions from gradient fields?," in *Intenational Conference on Computer Vision (ICCV)*, 2006.

144. T. Simchony, R. Chellappa, and M. Shao, "Direct analytical methods for solving poisson equations in computer vision problems," in *IEEE Trans. Pattern Anal. Machine Intell.*, 1990, vol. 12, pp. 435–446.

145. ISO/IEC JTC1/SC29/WG11, "Depth estimation reference software (ders) 4.0," M16605, July 2009.

146. "http://vision.middlebury.edu/flow/eval/," .

147. "http://gpu4vision.icg.tugraz.at/," .

148. Robert Cormack, "The computation of retinal disparity," *Perception and Psychophysics*, vol. 37, no. 2, pp. 176–178, 1985.

149. Tzung-Han Lin and Shang-Jen Hu, "Perceived depth analysis for view navigation of stereoscopic three-dimensional models," *Journal of Electronic Imaging*, vol. 23, no. 4, pp. 043014, 2014.

150. Dorin Comaniciu and Peter Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

151. "http://www.wisdom.weizmann.ac.il/ bagon/matlab.html," .

152. ITU-R Recommendation J.144 (Rev.1), "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," 2004.

153. A Ninassi, O Le Meur, P Le Callet, and D Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, pp. 253 – 265, April 2009.

154. Hosik Sohn, Yong Ju Jung, Seong il Lee, Hyun Wook Park, and Yong Man Ro, "Investigation of object thickness for visual discomfort prediction in stereoscopic images," in *Proceedings SPIE 8288, Stereoscopic Displays and Applications XXIII*, 2012, vol. 8288.

155. Satoshi Toyosawa and Takashi Kawai, "Measurement of perceived stereoscopic sensation through disparity metrics and compositions," in *Stereoscopic Displays and Applications XXV*, San Francisco, California, USA, 2014, vol. 9011.

156. Donghyun Kim, Dongbo Min, Juhyun Oh, Seonggyu Jeon, and Kwanghoon Sohn, "Depth map quality metric for three-dimensional video," in *Proceedings SPIE 7237, Stereoscopic Displays and Applications XX*, 2009.

157. ITU-T Recommendation P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2012.

158. Michael G. Ross and Aude Oliva, "Estimating perception of scene layout properties from global image features," *Journal of Vision*, vol. 10(1):2, pp. 1–25, 2010.

159. Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145175, 2001.

160. Lutz Goldmann, Touradj Ebrahimi, Pierre Lebreton, and Alexander Raake, "Towards a descriptive depth index for 3D content : measuring perspective depth cues," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, USA, 2012.

161. R.G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, april 2010, vol. 32, p. 722 732.

162. Cheng-Wei Chen and Yung-Yaw Chen, “Recovering depth from a single image using spectral energy of the defocused step edge gradient,” in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 1981–1984.

163. Pierre Lebreton, Alexander Raake, Marcus Barkowsky, and Patrick Le Callet, “Open perceptual binocular and monocular descriptors for stereoscopic 3D images and video characterization,” in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015.