Evaluating depth perception of 3D stereoscopic videos

Pierre Lebreton, Alexander Raake, Marcus Barkowsky, Patrick Le Callet

Evaluating depth perception of 3D stereoscopic videos

Pierre Lebreton[1,2], Alexander Raake[1]*Member,IEEE,*
Marcus Barkowsky[2],*Member,IEEE,* Patrick Le Callet[2]*Member,IEEE*

[1]Assessment of IP-Based Applications, Telekom Innovation Laboratories, TU Berlin
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
[2]LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes
Rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France

Abstract

3D video quality of experience (QoE) is a multidimensional problem; many factors contribute to the global rating like image quality, depth perception and visual discomfort. Due to this multidimensionality, it is proposed in this paper, that as a complement to assessing the quality degradation due to coding or transmission, the appropriateness of the non-distorted signal should be addressed. One important factor here is the depth information provided by the source sequences. From an application-perspective, the depth-characteristics of source content are of relevance for pre-validating whether the content is suitable for 3D video services. In addition, assessing the interplay between binocular and monocular depth features and depth perception are relevant topics for 3D video perception research. To achieve the evaluation of the suitability of 3D content, this paper describes both a subjective experiment and a new objective indicator to evaluate depth as one of the added values of 3D video.

Index Terms

3D Video, Depth evaluation, Objective model, Quality of Experience, Content characterization

I. INTRODUCTION

3D is a current trend for television. The contribution of adding 3D is stated by some to be at the same level as the transition from monochrome to color. The added value of 3D depends on the source material and the quality degradations due to coding [1] [2], transmission and the display device [3]. As a counterpart to the relative added value of 3D, a new factor significantly impacts the general quality of experience (QoE): the visual discomfort [4]. Even though one can say that visual discomfort could also be present with 2D video [5], with 3D videos this factor is particularly predominant in the construction of the global experience of the observers.

This paper targets the evaluation of the depth perception of the original source materials. This choice is motivated by the fact that before speaking of quality degradation due to compression and transmission, the source signals need to be considered, since different sources are not of equal quality. 3D video content materials are of different depth quality or comfort, and thus already without transmission the general QoE is highly content dependent. Having a tool for evaluating depth will provide some insight into the appropriateness of a given 3D video sequence, and into whether the 3D effect may deliver a clear added value for that sequence.

The perception of depth may be considered as an iterative process of the depth-cue dependent recognition of the different elements or objects contained in the scene, and of associating a specific position in depth to these objects or elements. The position in depth can be determined either by comparing the position of one object relative to the other, or directly using some knowledge about what is observed.

The paper is structured as follows: section two presents the different dimensions which could be considered when evaluating depth, and outlines approaches for depth evaluation reported in the literature. The third and fourth sections describe the set-up and results of a subjective experiment conducted to facilitate the understanding of how observers judge the depth quality in natural and synthetic video sequences. Section five introduces a new objective model for depth evaluation. Section six provides the model results on the test database and gives some information on its current limitations and possible improvements. Section seven concludes the paper.

II. The evaluation of depth

Two main axes can be investigated when evaluating depth: the depth perception and the depth quality. These are two distinct aspects: the depth perception is about the ability to understand the organization of the different elements composing the scene. Considering the fact that the depth understanding of the scene is mainly based on identifying relative object positions and spatial object features, this is usually denoted as “scene layout” [6]. The second aspect, the depth quality, depends on whether the depth information is realistic or plausible. If for example, a scene shows the so-called “cardboard effect” [7], the scene layout will be perceived correctly: the different planes can be distinguished and their relation in space can be understood. However, the depth quality will not be high, since the scene will likely appear abnormally layered.

A. Depth layout

There are many factors which contribute to the general understanding of the scene’s depth layout. These can be decomposed into two classes, monocular and binocular cues [8]. The monocular cues provide depth information using information from single views. Monocular cues can be further divided into two sub-classes: static and motion-based cues. Illustrations for the static cues are depicted in Figure 1: light and shading [9], relative size [10], interposition [10], blur [11], textural gradient [12], aerial perspective [10], and linear perspective [13]. Motion based cues are: motion parallax [14] and dynamic occlusion [10]. In addition to the monocular cues, the binocular vision provides the binocular depth cues. Here, stereopsis is considered as one of the most important depth cues. The pupils of the two human eyes are shifted by approximatively 6.5 cm, which causes each retinal image to provide a slightly different view of the same scene. The slight difference between views is called retinal disparity. The brain is able to combine these two views into a single 3D image in a process called stereopsis (see Figure 2). The general perception of the depth layout results from a combination of the different sources of depth information. Cutting and Vishton [6] studied the contribution of some of these depth cues to the general understanding of the scene layout, and provide depth discrimination functions for each of the studied depth cues as a function of the visualization distance (Figure 3). The chart in Figure 3 provides limits related with each depth cue. Some of them are always discriminative (like the occlusion), others like binocular disparity contribute only for a limited visualization distance range. For example, binocular disparity contributes to the depth discrimination for distances from 0-17m from the observer, and can also be used to estimate the magnitude of the depth separation within a smaller distances range, 0-2m [15]. These results give some insight into a possible pooling of the depth cues. However, further studies are still required to understand how the global depth impression is formed, since the experiments underlying the results in Figure 3 were focusing on the minimum offset necessary to perceive a depth difference (threshold detection), and not a weighting of their contributions to overall depth. Some results are available which give information on a possible weighting of different depth cue contributions. Proposals exist to reduce one depth cue, namely disparity, while emphasizing another depth cue, blur, in order to reduce visual discomfort for 3D reproduction without reducing depth perception [16] [17] [18].

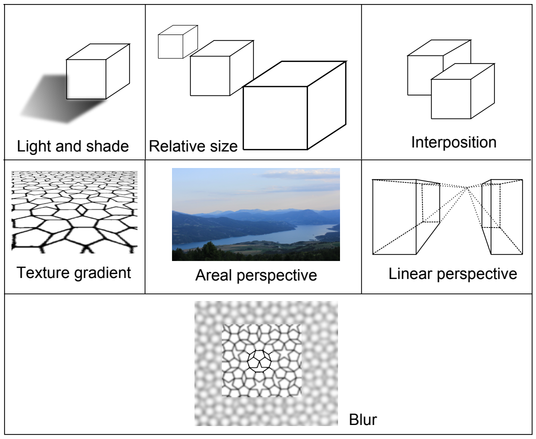

Figure 1: Different type of monocular cues

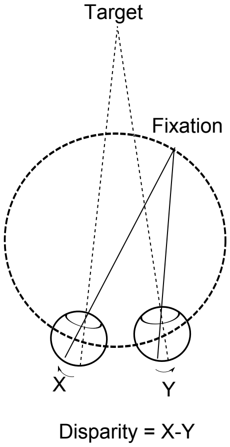

Figure 2: The retinal disparity used for stereopsis

B. Depth quality

The perceived depth quality is the other aspect of depth evaluation. Binocular disparity information has been added by stereoscopic 3D displays and may therefore be expected to play a major role when perceived depth quality is evaluated. Monocular and binocular depth cues may provide similar depth information, for example, it has been proposed to emphasize monocular depth cues in order to reduce visual discomfort from excessive disparity [17]. Many studies have been published on this topic, for example, in [19] the contribution of linear perspective to the perception of depth was analyzed. While work is still required on the interaction of the different monocular and binocular depth cues, this paper focuses mostly on the binocular depth cues since it plays a major role for discomfort and depth quality. Depth quality depends on different factors, and both the content and the display device will contribute to the general depth rating. As outlined above, the binocular cues contribute to the depth perception only within a limited distance range. To provide a good depth quality, it is required to have the objects of interest be positioned within this zone. 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 4). 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 [20] issues which need to be corrected in post-production. Both methods are valid, and there is no clear answer on which one to prefer. Once the zone where the camera converges is set (this is the null disparity plane), it is possible to adapt the distance between the cameras (e.g. the inter-camera distance) to ensure a good perception of the disparity cue. Figure 5 depicts how the depth quality is impacted by the inter-camera distance. Indeed, the shape of the voxels [3] (3D pixels) highly depends on the position of the cameras: if the inter-camera distance is too small compared to the distance of the cameras to the zero depth planes, the 3D images will appear layered (see Figure 5a). This is because the resolution in depth (illustrated by the voxels) is low; hence it is not possible to distinguish small variations of depth. Only high variations are resolved, and only a scene composed of discrete planes is visible. In turn, if the inter-camera distance is too high relative to the object distance, the resulting video may be perceived as uncomfortable due to too high values of disparities [4]. As a consequence, the entire depth budget is concentrated in a small depth zone (Figure 5b) , which however provides precise depth. Once the content is shot, the second factor for depth rendering is the display: displays also have limited depth resolutions (limited by display resolution). Hence, the depth is again quantized (Figure 6). For the present study, this last aspect will not be considered, and only the impact of the source sequence characteristics on depth perception of 3D video sequences will be addressed.

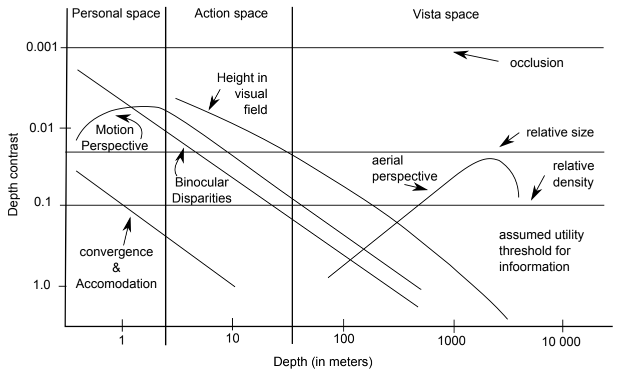

Figure 3: Depth contrast perception in function of the visualization distance. Results from [6]

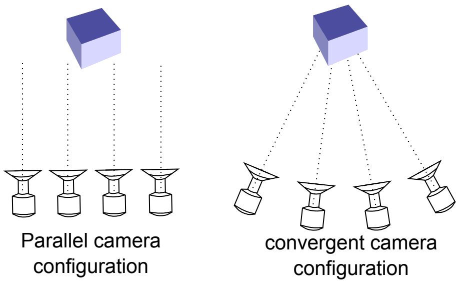
Parallel camera configuration
convergent camera configuration

Figure 4: Different type of camera's configuration

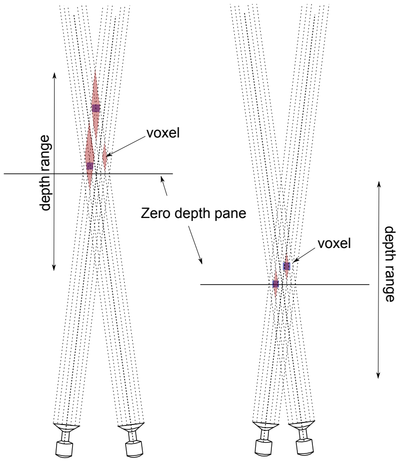

Figure 5: Effect of inter-camera distance on depth resolution considering the viewing distance

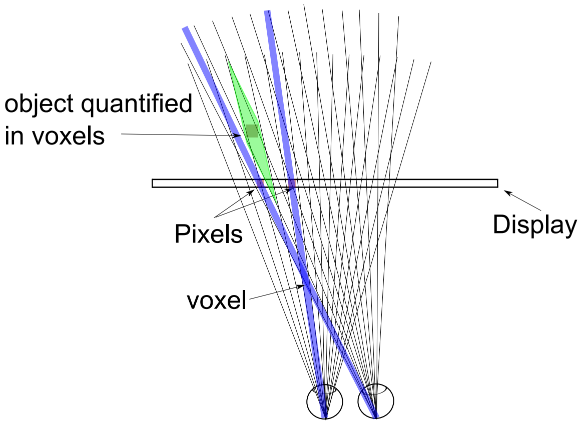

Figure 6: Effect display resolution on depth rendering

C. Depth evaluation results

Several studies reported in the literature target the evaluation of depth. Studies considering depth layout evaluation can use task-based experiments, with tasks such as the evaluation of the time required and the number of errors during the localization of a tumor in a brain [21], a depth ordering task, following a path in complex 3D graphs, or detect collisions [22], or evaluate the efficiency of air traffic control [23]. Other experiments like in Cuttin & Vishton [6] or [24] evaluate the visual discrimination abilities: two objects are presented in each trial, and the subject has to order the different objects. This step provides information on the depth contrast perception. Note that these experiments are typically performed using sythetic signals. Experiments evaluating depth in natural images are no direct alternative since the subject may be unable to distinguish between the depth quality and depth layout when asked to provide depth ratings. Oliva and al. [25] carried out subjective tests and evaluated the depth layout of still 2D images (considering only monocular cues). This, too, typically is a difficult task for the observers. Other interesting studies focused on the depth degradations due to compression [1]; here, the authors evaluated the quality, the depth and the naturalness of video sequences under different compression conditions, and relate quality with depth. The results are quality requirements to achieve a good depth perception. The spatial resolution and depth perception of stereoscopic images were evaluated by Stelmach et al. [2]. In their study, they observed that decreasing the horizontal resolution by half does not affect the depth perception too much. The effect of crosstalk on depth perception was studied in [26]. In the latter studies, it is difficult to determine the exact rating dimension of the subjects: depth layout or depth quality. They were asked for depth ratings, but without differentiating between depth layout and depth quality. It is likely that the test subjects provided ratings in terms of a combination of the two scales. These studies provide valuable input to the research presented here, but content-dependency of depth perception shall be first addressed.

D. Instrumental depth metrics

Objective metrics have already been presented for 3D QoE evaluation: quality metrics for stereoscopic 3D images are proposed in [27] [28] [29], and more specifically on depth evaluation, for the evaluation of the decrease of depth quality in [30] compared to a reference image or a similarity measure to evaluate the quality of depth maps in [31]. However these approaches assume that the reference has a perfect depth quality. Respective models extract a set of features to evaluate the depth quality degradation compared to the reference. For depth evaluation of source content, this approach is not meaningful, especially since no reference is available. In this case, a no-reference approach is required to evaluate content or depth quality of the source content.

III. SUBJECTIVE EXPERIMENT

To evaluate the depth, a database composed of 64 source reference signals (SRCs) has been designed. A description of each source sequence can be found in Table I. 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 [32]. 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.

IV. ANALYSIS OF RESULTS

The coherence of individual ratings of each observer with those of the other observers was checked by following the β_2 test as described in section 2.3.2 from ITU-R BT.500 [32]. The screening was done for each of the three

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

scales individually. Observer 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, 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. The results show a high correlation between the different scales (figure

7). 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 does 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 and eventually an influence of crosstalk between the two views although crosstalk was judged to be close to imperceptible by the experimenters. Previous studies [33] show that when observers are asked for quality evaluation, they usually do not take into account the depth in their rating (and then, 3D is not necessarily better rated than 2D). Indeed, in the experiment described by Seuntiens, subjects had to rate 3D still images which were shot with different inter-camera distances (0cm, 8cm, 12cm). These different inter-camera distances provide different binocular depth perceptions, but did not affect the image quality ratings significantly in [33]. Hence the added value due to depth did not seem to influence the quality ratings to a large extent. On the other hand, in presence of high disparity values as it may happen for sequences with a lot of depth, it may become more difficult for the observers to fuse the stereoscopic views [34] [35]. 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 [36]. 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 shooting and display conditions. As outlined earlier in this paper, an objective metric to quantify these differences will be useful for content suitability classification, and a first model algorithm will be presented in the following section.

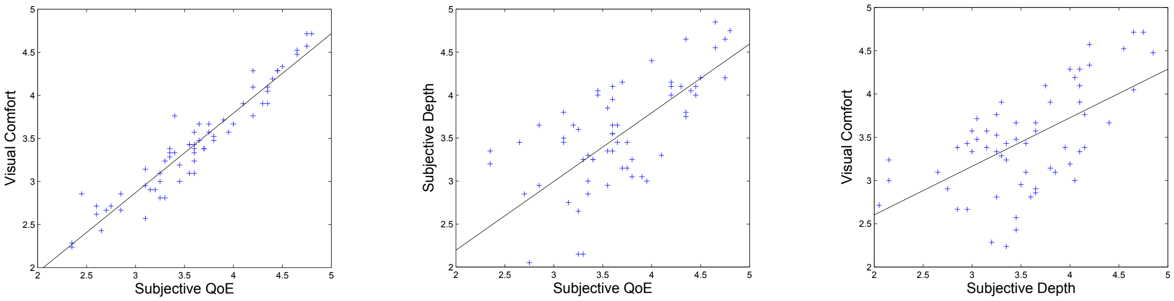

Figure 7: Scatterplots with regression lines showing the relation between the different evaluated scales

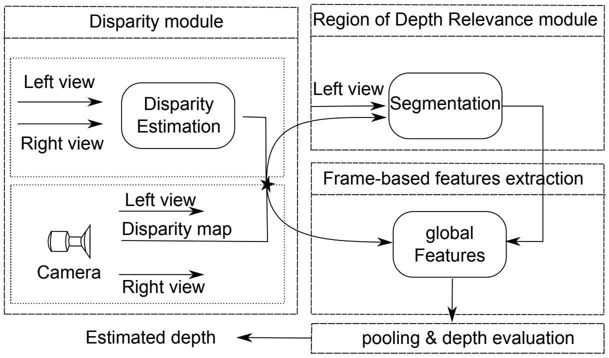

Figure 8: General structure of the proposed depth modelization

V. MODELLING

Considering that observers have watched dedicated 3D sequences in the subjective experiment presented here, and have also been asked to assess quality and visual discomfort in the same test, it is likely that they based their judgment of depth on the added value they perceived with the 3D stereoscopic representation. That is why even though monocular cues contribute to the depth perception, the prospective model presented here is based only on binocular cues. The general structure of the model is depicted in Figure 8. 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.

A. 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 multiples 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) [37] 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.

In the literature, studies establish the relation between disparity estimation and motion estimation. This is motivated by the analogy between the two tasks: finding pixel displacement between two frames. On the consecutive frames t and t+1 for motion estimation and on the stereo views for disparity estimation [38]. It has then been decided to use a dense optical flow algorithms to estimate the dense disparity maps. An extensive comparison of dense optical

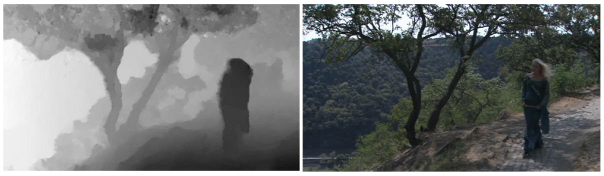

Figure 9: Results for the estimation of the disparity

flow algorithm is reported by the university of Middlebury [39]. Based on these results the algorithm proposed by Werlberger et al. [40] [41] and available at GPU4Vision [42] 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 9, and are obtained in a reasonable computation time (less than a second for processing a pair of full HD frames on an NVidia GTX470).

B. 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 10a, 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 10b 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 1 (also depicted in Figure 11):

Listing 1: Estimation of the region of depth relevance

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

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

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 introduced [43] [44]. This algorithm has been chosen due to its good performance in object segmentation on the data base under study, which has been verified informally for the segmented objects of a random selection of scenes.

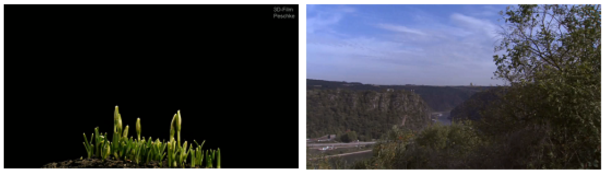

Figure 10: Illustration of cases where it is assumed that not the entire image is used for judging the depth.

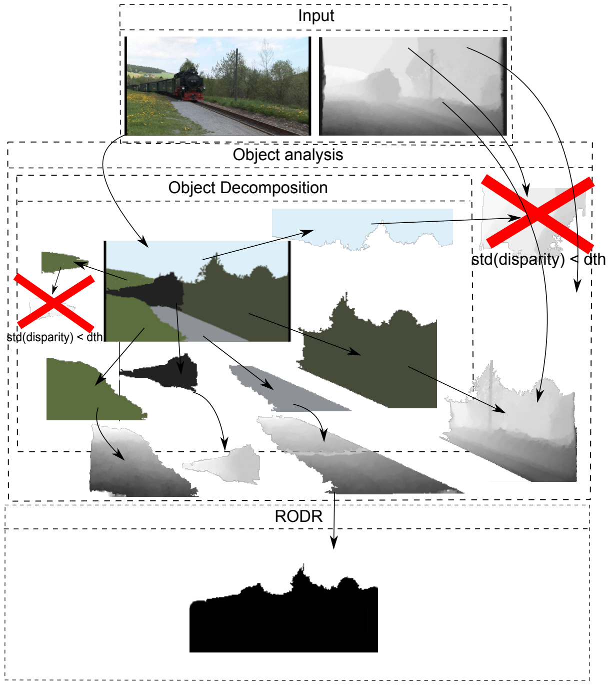

Figure 11: Illustration of the algorithm used for determining the region of depth relevance (RODR).

C. 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 2, as illustrated in Figure 12:

Listing 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.
```

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

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

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

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

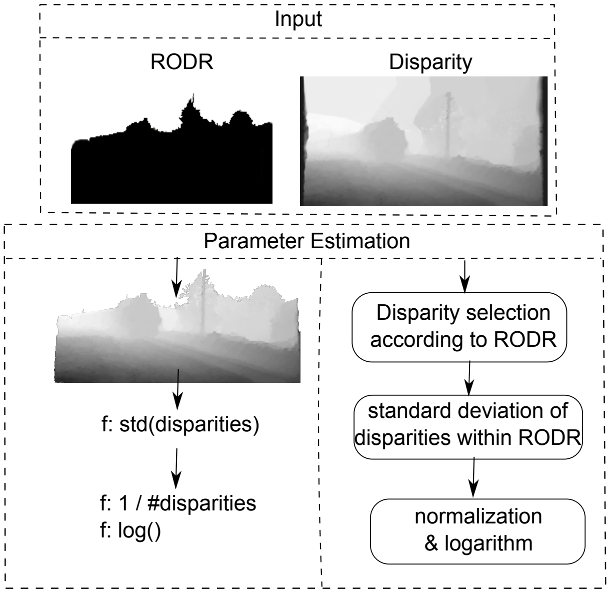
Input
RODR
Disparity
Parameter Estimation
f: std(disparities)
f: 1 / #disparities
f: log()
Disparity selection according to RODR
standard deviation of disparities within RODR
normalization & logarithm

Figure 12: Algorithm used for determining the value of the depth indicator for a single frame

D. Temporal pooling

Until now, no temporal properties of the 3D video sequences were considered. 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.

1) Short-term spatio-Temporal depth 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 [45], 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 3, and illustrated in Figure 13:

Listing 3: Spatio-Temporal metric

```
----------------------------------------------------------------------------------------------
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 depth indicator is defined as:

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


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

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

2) Global temporal pooling: Global temporal pooling is still work in progress: 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 [46] 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 [46]. 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

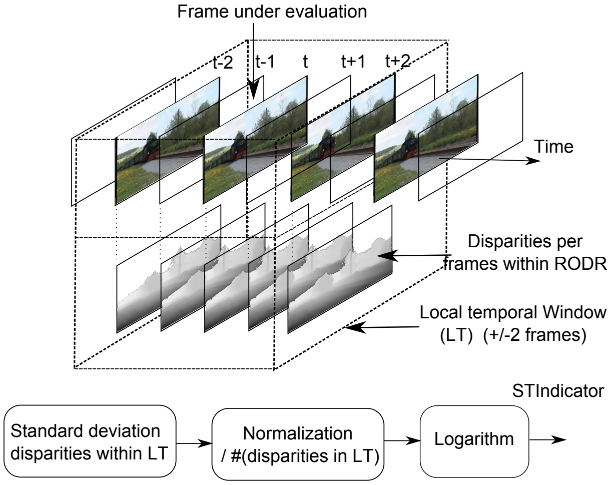
Frame under evaluation
t-2
t-1
t
t+1
t+2
Time
Disparities per frames within RODR
Local temporal Window (LT) (+/-2 frames)
STIndicator
Standard deviation disparities within LT
Normalization / #(disparities in LT)
Logarithm

Figure 13: Local temporal pooling

high short-term depth-values. The final mapping is then performed using a third order polynomial function.

Listing 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)

VI. Model performance

Figure 14 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 14, 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

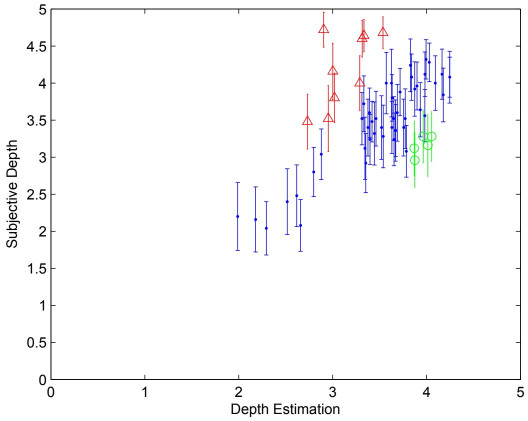

Figure 14: Results of the model on the estimation of depth, the triangles represents the contents which have pop-out effect, the circle represents a class of under-estimated content (which have a lot of linear perspective)

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 14), 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: 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 $X_{gth_i}$

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

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

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

VII. Conclusion

This paper describes an objective indicator for characterizing 3D materials on the depth perception scale. The model has been validated on a subjective depth evaluation database. The prediction performance of the model is promising even though several perception-related considerations are still missing, such as a weighting based on the position in depth of the object (if they pop-out or stay within the display). Further performance gain is expected with the improvement of the low-level feature extraction such as the disparity estimation, which shows its limits when it has to estimate the disparity of objects which pop out of the screen. Instead of an improved depth estimation algorithm, the depth recorded by cameras can be used in case of future set-ups to include this information during recording. The presented depth model forms the first part of a 3D material suitability evaluation framework. Further studies target the evaluation of other dimensions which play a role for the evaluation of the appropriateness of 3D video sequences.

References

[1] 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.

[2] 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.

[3] Chen Wei, Fournier Jérôme, Barkowsky Marcus, and Le Callet Patrick, "New requirements of subjective video quality assessment methodologies for 3DTV," in *Video Processing and Quality Metrics 2010 (VPQM)*, Scottsdale, USA, 2010.

[4] M. Lambooij, W. IJsselsteijn, and Heynderickx Fortuin, M., "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53(3), pp. 1–14, 2009.

[5] Eui Chul Lee, Kang Ryoung Park, Mincheol Whang, and Kyungha Min, "Measuring the degree of eyestrain caused by watching LCD and PDP devices," *International Journal of Industrial Ergonomics*, vol. 39, no. 5, pp. 798–806, September 2009.

[6] J. E Cutting and P. M Vishton, "Perceiving layout and knowing distance: The integration, relative potency and contextual use of different information about depth," in *Perception of Space and Motion*, S Rogers and W Epstein, Eds. 1995, pp. 69 – 117, New York: Academic Press.

[7] Hirokazu Yamanoue, Makoto Okui, and Fumio Okano, "Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images," in *IEEE Transations on circuits and systems for video technology*, 2006, vol. 16.

[8] Stephan Reichelt, Ralf Hussler, Gerald Ftterer, and Norbert Leister, "Depth cues in human visual perception and their realization in 3D displays," in *Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics IV*, Orlando, Florida, USA, April 2010.

[9] B. K. P Horn and M. J Brooks, "Shape from shading," *Cambridge: The MIT Press*, 1989.

[10] S. E Palmer, "Vision science: Photons to phenomenology," *Cambridge: The MIT Press.*, 1999.

[11] Vincent A. Nguyen, Ian P. Howard, and Robert S. Allison, "The contribution of image blur to depth perception," *Journal of Vision*, vol. 4, no. 8, 2004.

[12] B. J Super and A. C. Bovik, "Planar surface orientation from texture spatial frequencies," *Pattern Recognition*, vol. 28, pp. 729–743, 1995.

[13] A Criminisi, I Reid, and A Zisserman, “Single view metrology,” *International Journal of Computer Vision*, vol. 40, pp. 123–148, 2000.

[14] Ahmad Yoonessi and Curtis L. Baker Jr., “Contribution of motion parallax to segmentation and depth perception,” *Journal of Vision*, vol. 11, no. 9, pp. 1–21, 2011.

[15] Stephen Palmisano, Barbara Gillam, Donovan G. Govan, Robert S. Allison, and Julie M. Harris, “Stereoscopic perception of real depths at large distances,” *Journal of Vision*, vol. 10, no. 6, pp. 1–16, 2010.

[16] S. Watt, K. Akeley, and M. Ernst, “Focus cues affect perceived depth,” *Journal of Vision*, vol. 5, no. 5, 2005.

[17] Junle Wang, Marcus Barkowsky, Vincent Ricordel, and Patrick Le Callet, “Quantifying how the combination of blur and disparity affects the perceived depth,” in *Human Vision and Electronic Imaging XVI*, Thrasyvoulos N Rogowitz, Bernice EPappas, Ed. 2011, vol. 7865, pp. 78650K–78650K–10, Proceedings of the SPIE.

[18] George Mather and David R.R Smith, “Depth cue integration: stereopsis and image blur,” *Vision Research*, vol. 40, no. 25, pp. 3501–3506, January 2000.

[19] 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.

[20] T. Docherty A. Woods and R. Koch, “Image distortions in stereoscopic video systems,” in *Proceedings of the SPIE, Stereoscopic Displays and Applications IV*, 1993, vol. 1915, pp. 36–48.

[21] Jeremy R. Cooperstock and Guangyu Wang, “Stereoscopic display technologies, interaction paradigms, and rendering approaches for neurosurgical visualization,” in *Proceedings of SPIE, Stereoscopic Displays and Applications XX*, San Jose, CA, USA, 2009.

[22] Tovi Grossman and Ravin Balakrishnan, “An evaluation of depth perception on volumetric displays,” in *Proceedings of the working conference on Advanced visual interfaces, AVI*, New York, USA, 2006.

[23] Mark A. Brown, “On the evaluation of 3D display technologies for air traffic control,” Tech. Rep., Advanced Computing Environments (ACE) Laboratory, 1994.

[24] D.M. Hoffman, A.R. Girshick, K. Akeley, and M.S. Banks, “Vergence-accommodation conflicts hinder visual performance and cause visual fatigue,” *Journal of Vision*, vol. 8, no. 3, pp. 33, 2008.

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

[26] L.M.; Allison R.S. Tsirlin, I.; Wilcox, “The effect of crosstalk on the perceived depth from disparity and monocular occlusions,” *IEEE Transactions on Broadcasting*, vol. Volume 57, Issue 2, pp. 445 – 453, June 2011.

[27] 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.

[28] Sumei Li, Wei Xiang, Fuwei Cheng, Ruichao Zhao, and Chunping Hou, “HVS-based quality assessment metrics for 3D images,” in *Second WRI Global Congress on Intelligent Systems*, Wuhan, Hubei China, 2010, pp. 86 – 89.

[29] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, S Villette, and A.M. Kondoz, “Quality evaluation of color plus depth map-based stereoscopic video,” *IEEE journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304 – 318, April 2009.

[30] S.L.P. Yasakethu, D.V.S.X. De Silva, W.A.C. Fernando, and A. Kondoz, “Predicting sensation of depth in 3D video,” *Electronic Letters*, vol. 46, no. 12, pp. 837 – 839, June 2010.

[31] K. Wegner and O Stankiewicz, “Similarity measures for depth estimation,” in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Potsdam, Germany, 2009, pp. 1 – 4.

[32] ITU-R Recommendation BT.500-12, “Methodology for the subjective assessment of the quality of television pictures,” 2009.

[33] P Seuntiens, *Visual experience of 3D TV*, Ph.D. thesis, Eindhoven University, 2006.

[34] 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.

[35] 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.

[36] 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.

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

[38] Sourimant Gaël, "Depth maps estimation and use for 3DTV," Tech. Rep., INRIA Institut National de Recherche en Informatique et en Automatique, 2010.

[39] "http://vision.middlebury.edu/flow/eval/," .

[40] 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.

[41] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, June 2010.

[42] "http://gpu4vision.icg.tugraz.at/," .

[43] 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.

[44] "http://www.wisdom.weizmann.ac.il/ bagon/matlab.html," .

[45] 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.

[46] 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.

Pierre Lebreton received the engineering degree in computer science from Polytech'Nantes, Nantes, France after accomplishing an internship at NTT Service Integration Laboratories (Musashino, Tokyo, Japan) on the study of full reference metrics for assessing transcoded video quality in 2009. From January 2010 to June 2010 he worked as a research engineer at the Image and Video Communication (IVC) lab from CNRS IRCCyN on the study of video quality assessment of videos with different conditions of viewing. He is currently pursuing the Ph.D degree from the technical university of Berlin (TUB) in a join research within T-Labs and IVC. His research interests include 3D video quality evaluation, depth perception and understanding how the different 3D factors: pictorial quality, depth and comfort contribute to the general acceptance of 3D video sequences.

Alexander Raake is an Assistant Professor heading the group Assessment of IP-based Applications at Telekom Innovation Laboratories (T-Labs), Technische Universität (TU) Berlin. From 2005 to 2009, he was a senior scientist at the Quality & Usability Lab of T-Labs, TU Berlin. From 2004 to 2005, he was a Postdoctoral Researcher at LIMSI-CNRS in Orsay, France. From the Ruhr-Universitt Bochum, he obtained his doctoral degree (Dr.-Ing.) in January 2005, with a book on the speech quality of VoIP (extended version appeared as Speech Quality of VoIP, Wiley, 2006). After his graduation in 1997, he took up research at EPFL, Lausanne, Switzerland on ferroelectric thin films. Before, he studied Electrical Engineering in Aachen (RWTH) and Paris (ENST/Télécom). His research interests are in speech, audio and video transmission, Quality of Experience assessment, audiovisual and multimedia services and user perception modeling. Since 1999, he has been involved in the standardization activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals, where he currently acts as a Co-Rapporteur for question Q.14/12 on audiovisual quality models, and is co-editor of several standard recommendations. He has been awarded with a number of prices for his research, such as the Johann-Philipp-Reis-Preis in 2011.

Marcus Barkowsky received his Dipl.-Ing. degree in Electrical Engineering from the University of Erlangen-Nuremberg, Germany, in 1999. Starting from a deep knowledge of video coding algorithms his Ph.D. thesis focused on a reliable video quality measure for low bitrate scenarios. Special emphasis on mobile transmission led to the introduction of a visual quality measurement framework for combined spatio-temporal processing with special emphasis on the influence of transmission errors. He received the Dr.-Ing. degree from the University of Erlangen Nuremberg in 2009. Since November 2008, he is researching the relationship between the human visual system and the technological issues of 3D television at the University of Nantes, France. His current activities range from modeling the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays.

Patrick Le Callet received M.Sc. (1997) and PhD (2001) degree in image processing from Ecole polytechnique de luniversit'e de Nantes. He was also student at the Ecole Normale Superieure de Cachan where he get the Aggrgation (credentialing exam) in electronics of the French National Education (1996). He worked as an Assistant professor from 1997 to 1999 and as a full time lecturer from 1999 to 2003 at the department of Electrical engineering of Technical Institute of University of Nantes (IUT). Since 2003, he teaches at Ecole polytechnique de luniversit'e de Nantes (Engineer School) in the Electrical Engineering and the Computer Science department where he is now Full Professor. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. His current centers of interest are color and 3-D image perception, visual attention modeling, video and 3-D quality assessment. He is co-author of more than 150 publications and communications and co-inventor of 14 international patents on these topics. He has coordinated and is currently managing for IRCCyN several National or European collaborative research programs representing grants of more than 3 million Euros. He is serving in VQEG (Video Quality Expert Group) where is co-chairing the “Joint-Effort Group” and “3DTV” activities. He is the French national representative of the European COST action IC1003 QUALINET on Quality of Experience of Multimedia service in which he is leading the working group mechanisms of human perception.