Three factors that influence the overall quality of the stereoscopic 3D content: image quality, comfort, and realism

Raluca Vlad, Patricia Ladret, Anne Guérin-Dugué

Three factors that influence the overall quality of the stereoscopic 3D content: image quality, comfort and realism

Raluca Vlad, Patricia Ladret and Anne Guérin

GIPSA-lab, 11 rue des Mathématiques, Grenoble Campus,
BP46 F - 38402 Saint Martin d'Hères Cedex, Grenoble, France

ABSTRACT

In today's context, where 3D content is more abundant than ever and its acceptance by the public is probably definitive, there are many discussions on controlling and improving the *3D quality*. But what does this notion represent precisely? How can it be formalized and standardized? How can it be correctly evaluated? A great number of studies have investigated these matters and many interesting approaches have been proposed. Despite this, no universal 3D quality model has been accepted so far that would allow a uniform across studies assessment of the overall quality of 3D content, as it is perceived by the human observers.

In this paper, we are making a step forward in the development of a 3D quality model, by presenting the results of an exploratory study in which we started from the premise that the overall 3D perceived quality is a multidimensional concept that can be explained by the physical characteristics of the 3D content. We investigated the spontaneous impressions of the participants while watching varied 3D content, we analyzed the key notions that appeared in their discourse and identified correlations between their judgments and the characteristics of our database. The test proved to be rich in results. Among its conclusions, we consider of highest importance the fact that we could thus determine three different perceptual attributes – *image quality*, *comfort* and *realism* – that could constitute a first simplistic model for assessing the perceived 3D quality.

Keywords: 3D quality, stereoscopic quality, subjective evaluation, 3D database classification

1. INTRODUCTION

Two "waves" of commercial success for 3D have been noted in the past, one in the 1950s, and one in the 1980s. Both imposed themselves at first with attractive high-grossing movies, but both faded slowly away from public interest soon afterwards because of the low visual quality of the stereoscopic material of the time.

Right now, we are also in the middle of a new "3D wave", with the public strongly attracted by 3D content on various supports, from 3D cinema screens to personal phones with 3D capabilities. However, the step to overcome for the 3D to be generally "accepted" in the long term is considered to be the introduction of 3D in the homes and the gradual replacement of the 2D TV with the 3D TV, similarly to the historical transition from black and white TV to color TV in the 1960s and 1970s.

What is certain is that the introduction of 3D TV can only be a lasting success if the perceived image quality and the viewing comfort are at least comparable to conventional television. This is becoming increasingly feasible due to recent technological advances in image processing, camera and display development, as well as due to an enhanced understanding of 3D perception through human factor studies.[1] Still, what is lacking is a universal 3D image quality model, able to comprehensively define the concept of 3D quality as perceived by the user and able to standardize this concept in order to allow the correct comparison of 3D systems among experimenters.

The aim of the current paper is twofold. First, the paper aims to present the importance and the necessity of a standardized 3D image quality model. Second, it illustrates the implementation and the results of an exploratory study aimed to identify what could constitute the basis of a framework for evaluating the overall perceived quality of stereoscopic 3D images.

Further author information:
Contact e-mail: raluca-ioana.vlad@gipsa-lab.grenoble-inp.fr

The paper is structured as follows: in Sec. 2 we present the research context in which we are situating our approach, in Sec. 3 we give an overview of the subjective experiment that we implemented, in Sec. 4 and 5 we present in detail the implementation and the results of both parts of our study and we conclude with general observations and perspectives in Sec. 6.

2. THE IMAGE QUALITY CIRCLE OF ENGELDRUM AND THE CONCEPT OF IMAGE QUALITY MODEL

The Image Quality Circle (Figure 1) is a framework that relates the technology variables of an imaging system to the customer quality preferences and thus allows modeling the judgment of image quality independent of references.[2] This framework was first presented in the context of print quality,[3] but its generalized schema is suitable for modeling the quality percept for a large range of other systems involving visual content, including 3D systems.

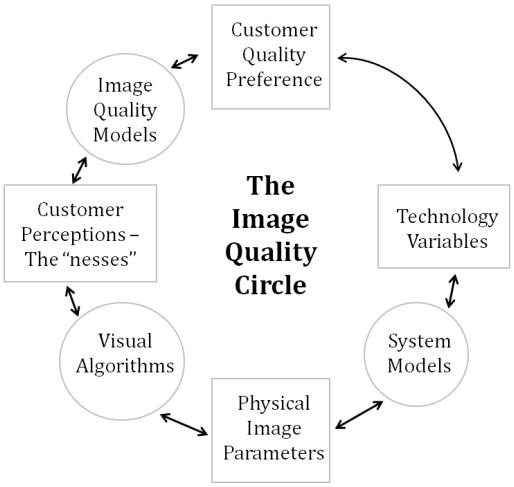


Figure 1. Engeldrum's Image Quality Circle

The focus in our paper will be on the concept of Image Quality Model, which is a key component of the Image Quality Circle. This concept, for a given system, can be defined formally by the perceptual attributes specific to that system and by the manner in which these attributes can be integrated to predict the image quality, in a similar way as they would be integrated by the human mind in an overall judgment.

To date, the most accurate ways to evaluate the perceived 3D quality are the subjective assessment methods, which use human evaluators that express their opinions on the content produced by a given system. The drawback is that these studies are time-consuming, meticulous and need to be repeated each time a parameter changes in the 3D system under study. Moreover, no universal framework exists yet that adapts the current 2D standards to be used correctly for the subjective assessment of the 3D content. At present, the most frequently used framework for evaluating 3D images and videos is the ITU-R BT.500 recommendation,[4] a standard designed for 2D content. Therefore, it is adapted differently by each experimenter that studies aspects related to 3D content, in function of their needs, thus leading to a much too large diversity in the experiment methodologies in order to allow universal terms of comparison among the studies performed around the world.

Having a standardized 3D image quality model would mean that the subjective assessments could be replaced by algorithmic solutions, which would be faster, more efficient, and also uniform across different studies on 3D perception.

One can still doubt if the existence of a model that predicts human perception from image characteristics is possible, but in an interesting experimental approach, Eerola et al. showed that indeed visual quality experience can be predicted by using measurable physical and computational level features.[5] This experiment concentrated

on print quality, but its generalized conclusions can stand true in any context in which visual quality needs to be assessed.

Thus, one of the current challenges in the context of 3D systems today is to elaborate such a 3D image quality model, by identifying the appropriate perceptual attributes that define *overall 3D quality* and by finding the influence of each one of them on the global percept. This is the challenge to which the current paper also tries to bring some answers.

In the search of the ideal 3D quality model, on one hand, there were numerous studies that tried to define the 3D stereoscopic quality only in relation to the 2D components of the stereoscopic content, both in the case of images and video.[6–8]

On the other hand, there were also studies in which the stereoscopic 3D quality was regarded as a much more complex multidimensional concept. For example, Meesters et al. enumerated factors like depth reproduction, stereoscopic impairments, and visual comfort to be taken into account when designing a model to evaluate the 3D quality.[1] Also, Seuntiens regarded the Image Quality Circle of Engeldrum[2] as a suitable framework for 3D quality assessment, but considered that the *image quality* concept cannot be applied as is to the 3D systems, and that a new concept, respectively *viewing experience*, would be more appropriate to include the added value of 3D content. He first considered the schema presented in Figure 2(a), and then refined it into the schema in Figure 2(b) as a model for evaluating the human visual percept when presented with 3D image.[9]

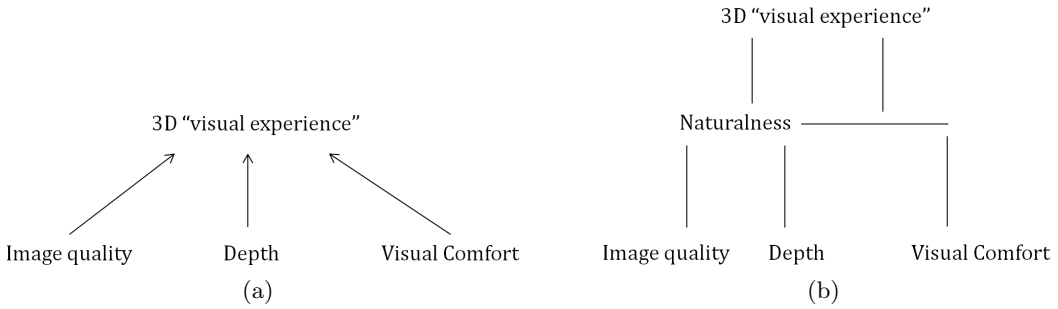


Figure 2. The 3D visual experience model of Seuntiens[9]

In our opinion, the former approach is incomplete in a context where the focus is on the overall 3D perceived quality, since, by concentrating only on the 2D aspects, the so-called *added value* of stereoscopic 3D (i.e. all the positive or negative effects of the represented depth on human perception) is not taken into consideration. An approach of this kind could however successfully model a part of what overall 3D quality means or it could be applied as is only in very specific contexts or studies, where the focus is strictly on the 2D quality of the composing views of the stereoscopic content. On the contrary, we consider the idea of multidimensionality appropriate to model the global human perception when watching 3D content and our study supported this concept and tried to define it through the means of a psycho-visual subjective experiment.

3. EXPERIMENT ON THE OVERALL 3D QUALITY PERCEIVED

When the bases of a new framework need to be set, the exploratory studies are essential. These are studies during which no direct subjective ratings are acquired, but instead unprimed attitudes, feelings and reactions towards the 3D content are explored.[1] The experiment that we present in this article was such an exploratory study that allows to arrive at a better understanding of the attributes underlying a multidimensional concept.[1] The notion that we examined was *overall 3D perceived quality*.

Despite the new terminology proposed by Seuntiens,[9] we preferred to adhere to the concept of *overall 3D perceived quality* during our experiment for referring to the general human percept triggered by watching 3D images. We consider that everything related to the 3D viewing experience can be encompassed by this concept.

3.1 Experiment structure

The exploratory experiment that we carried out consisted of two parts. The first part was a classification step, where the participants validated our subjective view on the structure of our 3D database and confirmed the desired variability of its content. The second part was a visualization step, where the viewers gave their verbal opinions on the perception of a selection of stereoscopic images they watched. The two parts were complemented by a testing stage before the visualization step, which allowed eliminating the participants with visual deficiencies, and also by a questionnaire at the end, which allowed gathering demographic data.

3.2 Participants

A total of 27 persons, aged between 21 and 43 (average age: 29) and mostly male (70%), participated in the experiment. The majority of the participants were highly qualified, mostly doctoral students, researchers and academics working in technical fields. None of the participants was directly involved in research related to 3D quality.

3.3 Vision tests

The evaluation of the visual acuity of the participants was done with a typical Snellen chart. Two persons had poor results at this test and, even if they participated in the visualization part of the experiment, their contribution was discarded in the analysis of the experiment data. The rest of the participants had a normal or corrected to normal view. For testing the color vision, the Ishihara test was used and all the participants proved a good color vision.

The stereo-acuity was assessed online, using a test made available by the McGill University.[10] The test was carried out using a stereoscopic display of a 3D capable computer. The test did not offer ratings of high precision for the stereo acuity values, but it allowed detecting very easily whether a participant was able to see stereoscopic content correctly or not. Among the 27 participants to the visualization part of our test, one person was unable to see stereoscopic images.

4. PART I - DATABASE VALIDATION

4.1 Test material

The initial test material was composed of a database of 158 stereoscopic images, all taken with a Fujifilm FinePix Real 3D W3 camera at its best quality configuration. Also, the minimal focal distance was imposed for every picture, for uniformity in the geometrical conditions of the shooting, given the fact that the 3D camera we used allowed varying the focal distance.

The content of the images in the database was meant to represent situations that ordinary people (non-specialist) would photograph if they owned a compact stereoscopic camera or a mobile phone with a stereoscopic camera. Thus, the content of the images was represented by everyday life sequences like objects, interiors, touristic sites, landscapes, people or animals.

For structuring the database in order to correlate the results of the visualization experiment to the features of the 3D data, three image characteristics were considered as the three independent variables to define our image collection: *complexity*, *depth interval* and *presence or absence of 2D depth cues*. The *depth interval* characteristic referred to the distance between the closest and the furthest object in the real photographed space. For each characteristic considered, several conditions were possible: for the first characteristic: *low complexity*, *average complexity*, *high complexity*; for the second characteristic: *small depth interval*, *average depth interval*, *large depth interval*; and for the third characteristic: *2D depth cues present*, *2D depth cues absent*. The intersection of these criteria led to a total of 18 categories of images, each category being defined by one condition of each variable.

4.2 Testing procedure

Because all the images had been photographed by only two persons, their structuring of the database according to the chosen criteria could be biased. This is why during the first part of our study we used subjective testing with multiple participants in order to confirm or invalidate the good distribution of the images in the 18 categories and to facilitate the subsequent selection of test material for the second and the most important part of our test.

The left views of all the 158 stereoscopic images in the database were printed on cardboard paper (approximately A5 in size). The 24 participants who took part in this test were offered the collection of printed images and were asked to classify them using a schema representing the database structure that was displayed at large scale on a table. For this, the participants would place each printed image on the table, in the case corresponding to the category chosen for it, as it can be seen in Figure 3.

Figure 3. Image of the experiments room during Part I of our study

4.3 Data interpretation and results

After the first analysis of the data, a satisfying rate of 77% was achieved for the classification, with 121 images well classified in the 18 categories and 37 images for which the classification was inconclusive.

However, an unexpected result showed up: the distribution of the images in the 18 categories was not balanced, with almost all the categories that corresponded to the *2D depth cues absent* condition being empty or with very few images. There are several explanations possible for these results: the imprecise description of the *presence or absence of 2D depth cues* characteristic to the participants, the incorrect or different comprehension of this characteristic among the participants or just the difficulty for the photographers of finding real-world situations where no 2D depth cues are found by a human judge. The description that we gave to the participants on the notion of *2D depth cues* was: "elements in a 2D image that help the observer mentally reconstitute the image in three-dimensions only after watching it in 2D (elements that offer information on the display of the objects in space, on distances, on dimensions, on depths etc.)". Therefore, we think that this unexpected result shows most of all the ability of the human system to elaborate a mental 3D representation of a scene in most cases, even for the images in which the experimenters considered that no 2D cues were present.

The unbalanced distribution of the images in the 18 categories led to the necessity of a new logic in the structuring of the database. The *presence or absence of 2D depth cues* characteristic was discarded and we only classified the images in function of the remaining valid criteria (*complexity* and *depth interval*). Following this classification in function of only one criterion at a time, high rates of success have been achieved: 97% for the complexity criterion (with 154 images well classified) and 98% in the case of the depth interval criterion (with 155 images well classified). Moreover, the distribution of the images in the 2 groups of 3 categories was balanced this time, as it can be observed in Table 1.

The classification thus obtained constituted a basis for the selection of the images for the second part of our study and also for the analyses to be made on its results, allowing the exploration of various correlations between the subjective appreciations on the 3D images and their two considered features, *complexity* and *depth interval*.

Table 1. The final distribution of all the database images per category.

Complexity			Depth interval		
Low	Average	High	Small	Average	Large
41 images	60 images	53 images	50 images	52 images	53 images

5. PART II - QUALITATIVE EXPLORATION OF SELECTED 3D CONTENT

5.1 Test material

Starting from the structured database that we obtained after the first part of the experiment, we selected 24 images as the test material for this second part of our study, following three criteria.

The first criterion of selection was to cover the largest range of semantic content types possible; this can be observed in Table 2.

Table 2. The 24 images selected for Part II of the experiment

Image 1	Image 2	Image 3	Image 4	Image 5	Image 6
					
High complexity Average depth interval	High complexity Small depth interval	High complexity –	High complexity Average depth interval	High complexity Small depth interval	High complexity Average depth interval
Image 7	**Image 8**	**Image 9**	**Image 10**	**Image 11**	**Image 12**
					
High complexity Average depth interval	Average complexity Large depth interval	High complexity Large depth interval	Low complexity Large depth interval	Average complexity Average depth interval	High complexity Small depth interval
Image 13	**Image 14**	**Image 15**	**Image 16**	**Image 17**	**Image 18**
					
High complexity Small depth interval	Average complexity Average depth interval	Average complexity Large depth interval	Average complexity Small depth interval	Average complexity Small depth interval	Low complexity Small depth interval
Image 19	**Image 20**	**Image 21**	**Image 22**	**Image 23**	**Image 24**
					
Low complexity Average depth interval	Average complexity Small depth interval	Low complexity Small depth interval	Average complexity Average depth interval	High complexity Large depth interval	High complexity Large depth interval

The second criterion of selection was to cover all the 6 categories previously defined; we considered that a minimum of 4 images per category was sufficient. The distribution of the 24 images according to the complexity and real depth interval is given in Table 3 and the categories to which the images belong can be found in Table 2.

Table 3. The final distribution of the database images per category

Complexity			Depth interval		
Low	Average	High	Small	Average	Large
4 images	8 images	12 images	9 images	8 images	6 images

The third criterion of selection was related to the depths represented on screen, which are given by the disparities that are present in the stereoscopic couples. A stereo-matching algorithm implementing the SIFT feature detection method[11] was used for computing the disparities of a large number of feature points in each of the 24 images. This allowed us to determine their disparity ranges and to select images of varying represented depths for the Part II of our study. The graphical representation of the horizontal disparity ranges of each of the 24 images, given in pixels for images of size $3584x2016$, is in Figure 4.

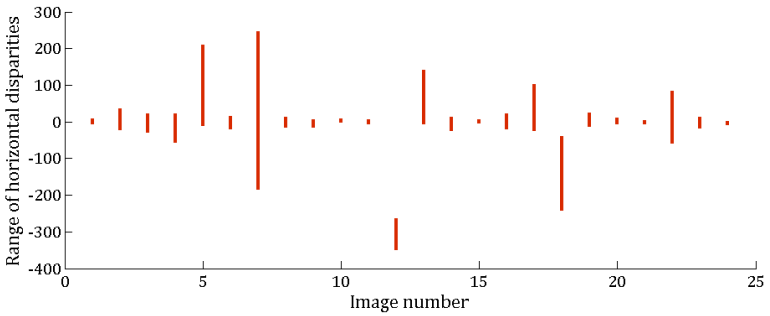


Figure 4. The horizontal disparity ranges for the 24 images

5.2 Testing procedure

The visualization experiment was carried out with a number of 26 participants and only the results of the 24 participants with a good visual condition were considered for analysis in the end.

For the visualization experiment, a Panasonic TC-P50VT20 stereoscopic display with active shuttering glasses was used. The television screen was situated on a table in front of the participant, who was advised to change the seat height for an optimal position, centered relative to the screen. The visualization distance was of 140 cm and it was chosen as a compromise between the ITU Recommendation for assessing stereoscopic pictures[4] and the television specifications. A daylight lamp of 25W was used to light the room of approximately 12 m^2 in surface.

The qualitative exploration of the 3D perceived quality was carried out in the form of a session of visualization of the series of 24 3D images, during which the participants were supposed to freely express their opinions on the perceived quality and on the comfort/discomfort induced by the 3D content.

The experiment was a qualitative exploration[12] and followed the principles of a semi-structured interview.[13,14] The same explanatory text was offered to all the participants, enumerating the two main dimensions on which they were supposed to express their opinions, respectively *quality* and *comfort*. After each participant finished verbalizing his or her first impressions, extra questions were asked in order to clarify these first impressions and to make sure that the factors that were possibly considered irrelevant by the test subjects were also discussed.[13] The supporting questions allowed as well maintaining the focus of the observer on elements directly related to the overall 3D perceived quality and avoiding comments unrelated to the purpose of the experiment.

No time limit was set for the verbalization experiment; the participants were free to speak for as long as they wanted about how they perceived a given 3D image. The average visualization and verbalization time per participant for the totality of 24 images was of 28 minutes.

The test was held either in English or French, in function of the preferred language of the participants, and the remarks that were made were recorded in ordered to be listened and processed later.

5.3 Data interpretation and results

All the participant remarks were extracted from the audio recordings and organized in a 24 x 24 observations table, in which each case represented the observations made by one participant (or "judge") on one image.

The characteristics on which the participants had focused the most were extracted from the text. The *quality* and the *comfort* characteristics stood out as expected. But it turned out that the participants had also shown much importance to the idea of *realism* and to a particular artifact, the *cardboard effect* – distortion generated by the disproportionate scaling of the real depths into the depths represented on screen –, and had made detailed observations related to them.

These four characteristics were subsequently studied in detail by extracting key words referring to each of them from the text table. Based on the key words, numeric ratings were associated to each case of the text table for each of the characteristic in relation to specific scales. The choice of the complexity and type of the scale for each characteristic was determined by the degree of detail that was present in the comments made by the participants or by the nature of the characteristic itself. Five-rating scales were chosen to represent the data for quality and comfort, a three-rating scale for realism, and a binary scale for the cardboard effect, as in Figure 5. For quality, the five possible ratings represented the following tags: *very low quality*, *low quality*, *neutral*, *high quality* and *very high quality*. Similarly, the ratings for comfort corresponded to: *very uncomfortable*, *uncomfortable*, *neutral*, *comfortable* and *very comfortable*. The comments made on the realism were coded by: *artificial*, *neutral* or *realistic*. As for the cardboard effect, the choice was binary, evaluating the presence or the absence of this artifact.

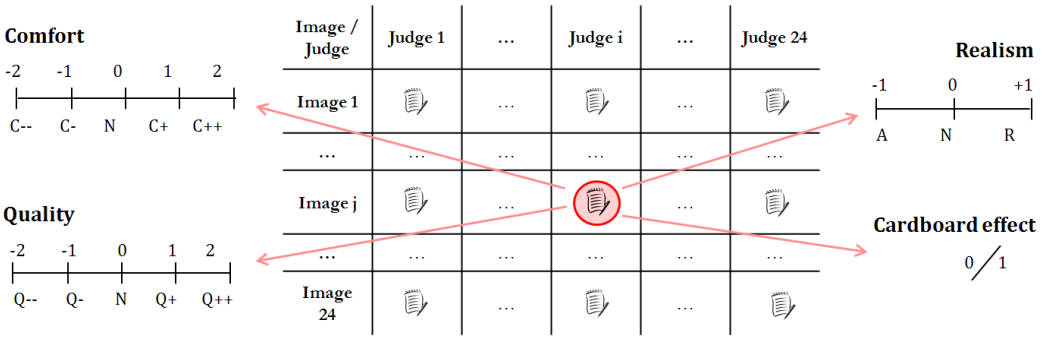


Figure 5. The rating scales used in the text analysis

In order to illustrate how the association between key words and ratings has been made, in Table 4 there are a few examples, where each rating is in brackets, after the group of words to which it was attributed.

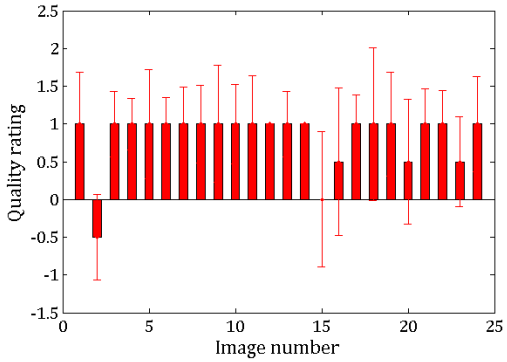
All the subsequent studies were made by analyzing these ratings. The method used for computing an overall rating per image from the quality and comfort ratings was scales aggregation, in the form of the median value of all the votes for that image.[15] For realism, we used the adapted formula:

$$realism\ rating\ per\ image = (noP - noN)^2 * sgn\,(noP - noN) \,/\, (noP + noN)\,, \tag{1}$$

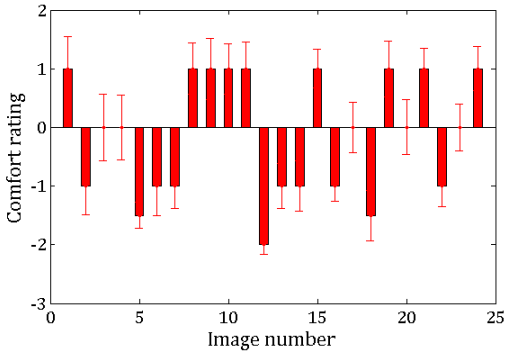
where noP and noN represent the number of positive votes and the number of negative votes for that image. And in the case of the cardboard effect, the rating per image was given directly by the number of participants that made comments related to the presence of this artifact (situations marked with '1' in the numeric table). The ratings obtained can be observed in Figure 6.

Table 4. Examples of extracted key words and their associated ratings

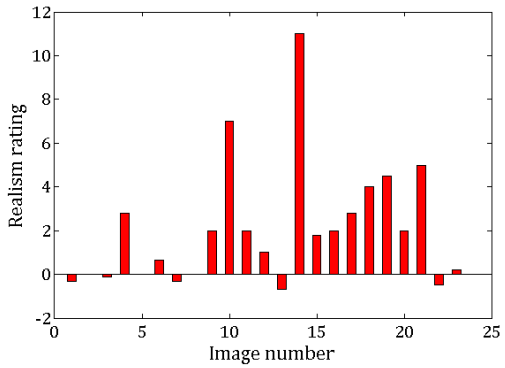
Quality	Comfort	Realism	Cardboard effect
very bad quality (-2)	*really annoying, uncomfortable* (-2)	*it's super artificial* (-1)	*impression of successive layers* (1)
the quality is below average (-1)	*a little bit annoying* (-1)	*it's not like in reality* (-1)	*no continuum in depth* (1)
average quality (0)	*I don't feel annoyed, but it's not comfortable either* (0)	*like in real life* (1)	*as if each object was flat* (1)
it's rather a good quality image (+1)	*generally speaking, it's comfortable* (1)	*natural* (1)	*a little like made of cardboard* (1)
very, very good quality image (+2)	*very relaxing* (2)	*I see myself projected in the image* (1)	*rather like having three planes, than a volume* (1)



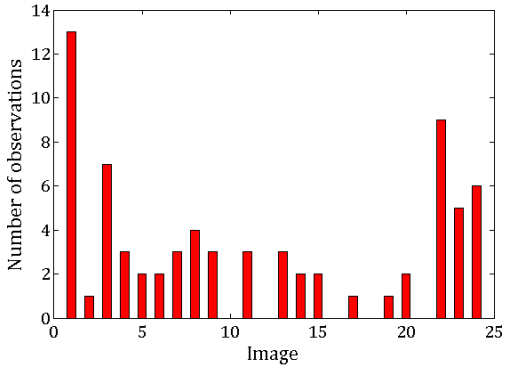
(a) The subjective quality ratings per image



(b) The subjective comfort ratings per image



(c) The subjective realism ratings per image



(d) The subjective cardboard effect ratings per image

Figure 6. The subjective ratings per image

5.4 Results on quality

The fact that almost all the stereoscopic images were judged as of *high quality* by the participants (see Figure 6(a)) proves that the notion of *quality* was an ambiguous term during the experiment. The participants have been asked to express their opinions on the *quality* of the stereoscopic images and the prior expectations were that they would include in this notion the totality of the aspects that contribute to the overall 3D experience. On the contrary, the responses provided by the participants on the notion of *quality* were proven to refer mostly

to the 2D quality of the views composing the stereoscopic image. This behavior can be explained by the fact that the 2D quality was very similar throughout the database, since all the images have been taken with the same camera in the same conditions, and the only images that obtained a lower quality score were those that contained some kind of 2D degradation (motion blur or elements that were perceived as lack of contrast).

Our conclusion is that the concept of *quality* should always be defined during a subjective experiment in function of the context. In our case, where we explore the perceptions triggered by 3D content, we consider that *image quality* can successfully represent a dimension of the notion of *overall 3D perceived quality*, but that the *image quality* concept should be related exclusively to the quality of the 2D views that make up the stereoscopic data.

5.5 Results on comfort

The subjective ratings on comfort were first analyzed in relation to the complexity categorization of the 24 images. We could observe that more images rated as uncomfortable were in the *high complexity* group and only one image judged as uncomfortable was in the *low complexity* group, this indicating a possible influence of complexity on comfort (Figure 7). However, the results of the statistical analysis using the one-factor Anova test and of the multiple comparison test show that there are no significant differences among the complexity categories for our comfort ratings ($p\_value = 0.4276$). From these statistical results we can conclude that an influence of image complexity on comfort could not be validated in the context of our experiment.

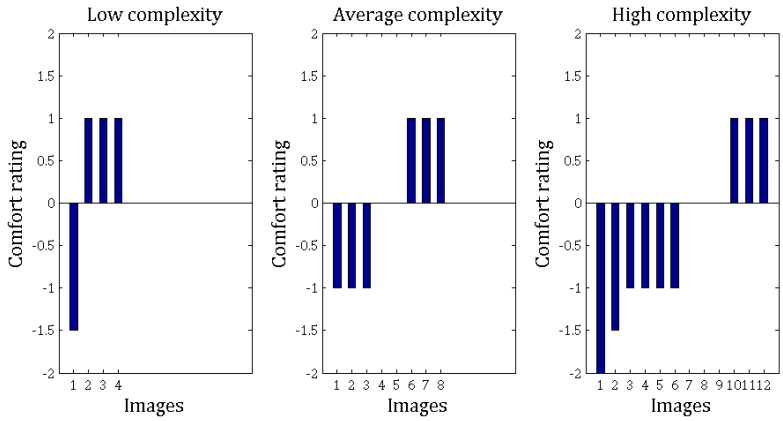


Figure 7. The subjective comfort ratings grouped in function of the complexity of the corresponding images

When grouping the 24 comfort ratings in function of the real depth intervals of the corresponding images, one category stands out as significantly different, compared to the others. The category of images with large depth intervals contains only images with positive comfort ratings, while in the other two categories there are mostly images with negative comfort ratings (Figure 8). The fact that large depth intervals are related to positive comfort in the context of our database is confirmed by the Anova test ($p\_value = 0.0074$) and by the multiple comparison test. This result shows that the comfort perceived can be influenced by the real depths of the photographed scene. Nevertheless, an overall correlation across all the three categories could not be validated in our study.

A third test related to comfort was done in order to verify the relation between the comfort ratings and the horizontal disparities of the 24 images displayed during Part II of the experiment. The correlation index of -0.5716 is sufficiently significant to indicate an inverse correlation between the two. The graphical representation, where the results per image are ordered in function of the comfort ratings (Figure 9), also shows that for small disparity ranges the comfort was usually positive, while for large disparity ranges the comfort ratings were always negative in our experiment. Thus, the results obtained indicate a negative influence of large disparity ranges on comfort and these results are consistent with our expectations, confirming that images with large represented depths need more effort for the fusion of their content, thus implying more visual discomfort.

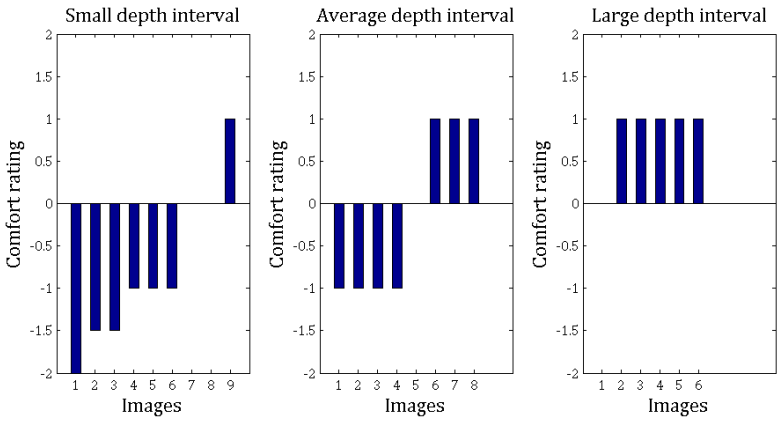


Figure 8. The comfort ratings grouped in function of the depth interval of the corresponding images

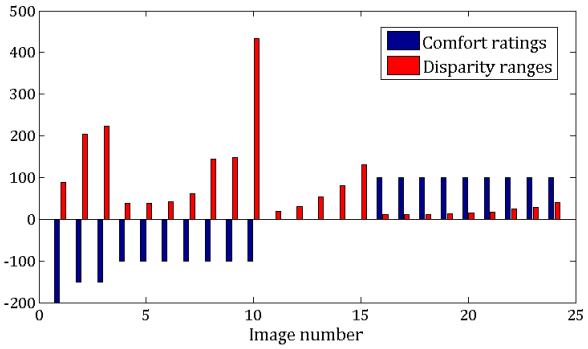


Figure 9. The comfort ratings and the disparity ranges per image, ordered in function of the comfort

Following the numerical analysis of the data on comfort, our conclusion is that this perceptual attribute is mostly influenced by the horizontal disparities that are present in the 3D images. Other characteristics of the 3D content can have an influence on comfort as well, but further narrower studies are necessary for an in-depth view on this matter.

5.6 Results on realism

For judging the realism percept in relation to the complexity of the 3D images, the 24 realism ratings were classified in function of the three complexity categories (Figure 10). The median values of the three groups of ratings were determined as distinct and the results of the statistical tests showed significant differences among them ($p\_value = 0.0043$). Thus, complexity appears to be influencing realism in the sense that images of low complexity have been considered more realistic and images of high complexity have been considered more artificial during our visualization test.

In the case of the real depth intervals, neither the graphical representation nor the statistical results ($p\_value = 0.8739$) could give indications on a correlation between this characteristic and the realism perceived (Figure 11).

The correlation test on the realism and disparity range data ($correlation\ index = 0.0706$) was not relevant either, suggesting that the disparities in the stereoscopic images did not influence the realism percept induced. Our test hypothesis, which supposed that the realism could be influenced by the presence of negative disparities that give a sensation of immersion, was invalidated.

A correlation index of -0.4689 between the realism ratings and the cardboard effect ratings suggested that the presence of the cardboard effect made the stereoscopic images seem more artificial. The conclusion is equally illustrated by the graphical representation of realism and cardboard effect in parallel (Figure 12), where it can

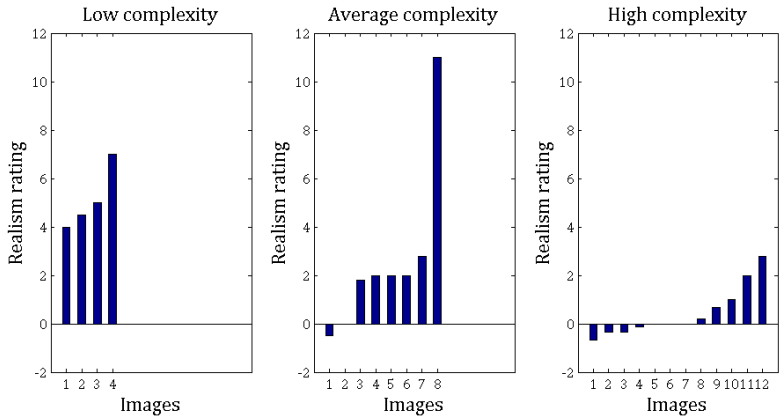


Figure 10. The realism ratings grouped in function of the complexity of the corresponding images

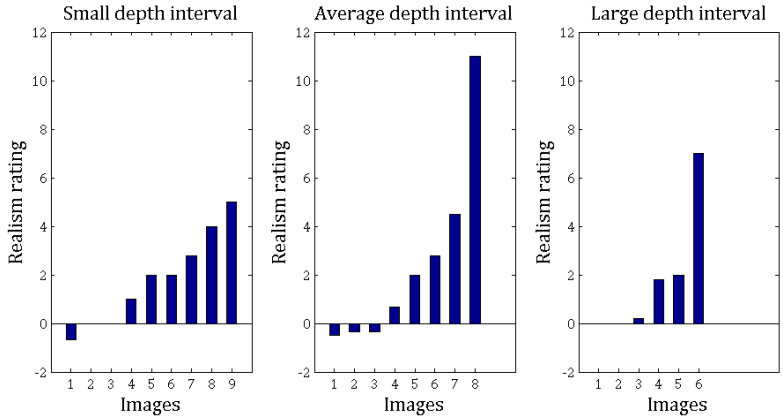


Figure 11. The realism ratings grouped in function of the depth interval of the corresponding images

be observed that for the images where the cardboard effect was pronounced, the ratings on realism were negative or neutral, suggesting a sensation of artificiality or no percept of realism.

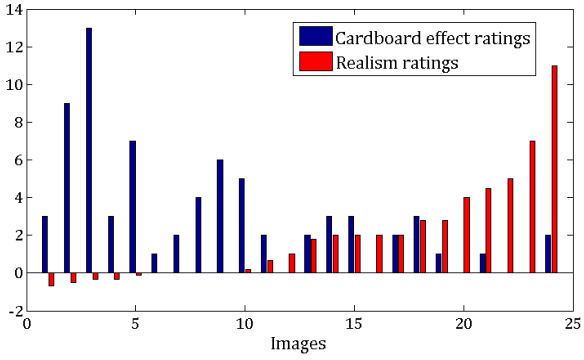


Figure 12. The cardboard effect ratings and the realism ratings per image, ordered in function of the realism

The observations that we could make related to the realism of 3D images encourage us to believe that among the tested factors, the cardboard effect had the strongest negative influence. The fact that high complexity images also triggered a sensation of artificiality for the images of our database can also be associated to the cardboard effect, which is better perceived in images with a larger number of elements.

Since the cardboard effect is a geometric distortion, we are also inclined to believe that other geometrical distortions could influence as well the percept of realism, but this hypothesis needs to be submitted to test.

5.7 Overall results of our exploratory study

Since what we are searching for is a basis for a 3D quality model, we tested the interactions between the three perceptual attributes that we could identify during the exploratory study: *quality* (considered by us to be actually the attribute *2D image quality*), *comfort* and *realism*. Thus, the correlation coefficients calculated for the subjective ratings on quality and comfort, quality and realism, and comfort and realism (0.0451, 0.1364, and 0.0931) clearly showed the lack of dependency between these three factors. This is a sufficient indicator of the fact that all of these three perceptual attributes could be included in an independent manner in the elaboration of a model that evaluates the overall 3D perceived quality (Figure 13).

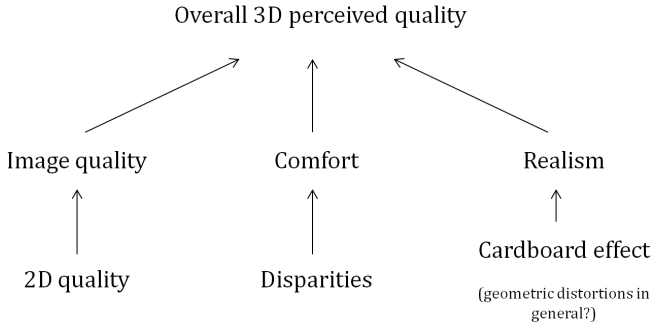


Figure 13. The proposed 3D quality model

6. CONCLUSIONS AND PERSPECTIVES

The purpose of our study was to underline the need for a universal 3D quality model and to present the exploratory approach that we used for determining which perceptual attributes could lay at its basis.

Our exploratory test proved to be rich in results. Among its conclusions, we consider of highest importance the fact that, starting from a database of varied and controlled content, we could extract three different perceptual attributes – *image quality*, *comfort* and *realism* – that influence the *overall perceived quality of stereoscopic 3D images*. Since we could identify the *2D quality*, the *horizontal disparities* and the *cardboard effect* as important physical properties of the 3D data that influence the three perceptual attributes mentioned, we consider that our results could be a starting point for the elaboration of a 3D quality model, by integrating all these dependencies in the Image Quality Circle of Engeldrum.

Our next research projects will be focused on validating the results of this first exploratory study in a more precise context, by looking into each of the three factors in detail and by trying to integrate them in an overall view on the perception of 3D content.

ACKNOWLEDGMENTS

The work presented in this paper was supported by the MOOV3D project, via the MINALOGIC competitive cluster, IMAGINOVE and Rhônes-Alpes region in France.

REFERENCES

[1] Meesters, L., IJsselsteijn, W., and Seuntiens, P., "A Survey of Perceptual Evaluations and Requirements of Three-Dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology* **14**(3), 381–391 (2004).

[2] Engeldrum, P., "Image Quality Modeling : Where Are We ?," in [*IS&T's PICS Conference*], 251–255 (1999).

[3] Engeldrum, P., "Measuring customer perception of print quality," in [*IS&T 42nd Annual Meeting*], 161–164 (1989).

[4] "Recommendation, ITU BT.500-13 - Methodology for the subjective assessment of the quality of television pictures," Recommendation, ITU (2012).

[5] Eerola, T., Kamarainen, J. K., Leisti, T., Halonen, R., Lensu, L., Kalviainen, H., Nyman, G., and Oittinen, P., "Is there hope for predicting human visual quality experience?," in [*IEEE International Conference on Systems Man and Cybernetics*], 725–732 (2008).

[6] Yasakethu, S., Fernando, W., Kamolrat, B., and Kondoz, A., "Analyzing perceptual attributes of 3D video," *IEEE Transactions on Consumer Electronics* **55**(2), 864–872 (2009).

[7] Joveluro, P., Malekmohamadi, H., Fernando, W., and Kondoz, A., "Perceptual Video Quality Metric for 3D video quality assessment," in [*3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*], 1–4 (2010).

[8] Campisi, P., Le Callet, P., and Marini, E., "Stereoscopic Images Quality Assessment," in [*European Signal Processing Conference (EUSIPCO)*], (2007).

[9] Seuntiens, P., *Visual Experience of 3D TV*, PhD thesis, Eindhoven: Technische Universiteit Eindhoven (2006).

[10] Hess, R. and Cooperstock, J., "Test your stereo (3D) vision." `http://3d.mcgill.ca/` (2012).

[11] Lowe, D., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision* **60**(2), 91–110 (2004).

[12] Jumisko-Pyykko, S. and Strohmeier, D., "Report on research methodologies for the experiments," Technical report, Mobile3DTV.

[13] Jumisko-Pyykko, S., Hakkinen, J., and Nyman, G., "Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality," *Multimedia on Mobile Devices, SPIE* **6507** (2007).

[14] Jumisko-Pyykko, S., Reiter, U., and Weigel, C., "Produced Quality is Not Perceived Quality - A Qualitative Approach to Overall Audiovisual Quality," in [*3DTV Conference, 2007*], 1–4 (2007).

[15] Marcotorchino, J. and Michaud, P., [*Optimisation en analyse ordinale des données Volume 4 of Statistique et décisions économiques*], Masson, Paris (1979).