TOWARDS A DESCRIPTIVE DEPTH INDEX FOR 3D CONTENT: MEASURING PERSPECTIVE DEPTH CUES

Lutz Goldmann, Touraj Ebrahimi, Pierre Lebreton, Alexander Raake

TOWARDS A DESCRIPTIVE DEPTH INDEX FOR 3D CONTENT: MEASURING PERSPECTIVE DEPTH CUES

Lutz Goldmann, Touradj Ebrahimi

Ecole Polyt. Federale de Lausanne
Multimedia Signal Processing Group
Lausanne, Switzerland

Pierre Lebreton, Alexander Raake

Telekom Innovation Labs & TU Berlin
Assessment of IP-Based Applications
Berlin, Germany

ABSTRACT

3D quality of experience (QoE) in nature is a multidimensional problem and involves many factors that contribute to the global quality rating such as image quality, depth perception and visual discomfort. One important aspect for the development and evaluation of 3D processing techniques is the selection of appropriate 3D content. To this aim it is necessary to develop computational methods that can automatically measure the 3D characteristics of a scene, similar to the spatial and temporal information indices commonly used for assessing 2D content. The presented work is one step in the development of such a depth index (DI) which will target the evaluation of the depth-related characteristics of 3D video sequences. The paper focuses on the linear perspective as one of the major monocular depth cues. It compares two distinct approaches for measuring the strength of perspective depth cues and analyzes their limits on a 2D image dataset with associated subjective ratings.

1. INTRODUCTION

Recently, the interest in 3DTV, as one of the emerging multimedia formats, has remarkably increased due to the rapid technological development.As a consequence efforts have been devoted to the creation of suitable 3D content to feed this growing market. However, one of the most important factors for a sustainable success of 3DTV is that it provides an increased quality of experience (QoE) when compared to traditional 2D media formats. Achieving this is not straightforward due to the various quality factors which are involved in a 3D experience. Three dimensions are usually considered when studying 3D quality: image quality, depth perception and visual discomfort.

In this paper we focus on the depth perception since this directly reflects the added value of 3DTV. There are various factors which contribute to the human depth perception. These depth cues can be divided into monocular (pictorial) and binocular cues [1]. The monocular depth cues provide depth information available in single views and include shading, relative size, interposition, blur, texture gradients, linear perspective, motion parallax and dynamic occlusions. In addition to the monocular cues, the binocular vision (stereopsis) provides binocular depth cues. Here, binocular disparity is considered as one of the most important binocular depth cues.

In this paper we focus on the contribution of monocular depth cues to the overall depth perception. As one of the strongest monocular depth cues the linear perspective has been used for many years by artists to give a good understanding of the depth layout of a scene. Another aspect which makes linear perspective particularly interesting is its ability to give quantitative depth information: using linear perspective it is possible to estimate the distance in depth between objects. This is not necessarily the case with other depth cues: if interposition is considered without additional information, it is only possible to say that one object is in front of another, but not at which distance. The linear perspective has been studied in the past and typically in two different granularities: firstly as a local depth measure [2, 3] where the linear perspective

Pierre Lebreton is also affiliated with the University of Nantes. The authors thank Marcus Barkowsky and Patrick Le Callet from the University of Nantes for their comments and fruitful discussions. This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET (http://www.qualinet.eu/).

is used to estimate a dense depth map of a scene, and secondly in a global way [4] where a single depth indicator is computed to provide information about the perceived depth. Both approaches are valuable but target different objectives. The first one could be used to understand more precisely the structure of the scene and could then be applied to improve the quality of depth maps or to evaluate distances to objects in the context of autonomous robots. The second approach directly targets the description of a scene, which can be particularly useful for the development of realiable 3D processing techniques, where it is crucial to select representative 3D content with different characteristics. While for 2D content, the spatial (SI) and temporal information (TI) indices are commonly used [5], an additional depth index (DI) is required, to capture the depth characteristics as well.

The general objective of our work is to develop a depth index which describes the depth characteristics of a scene by analyzing and fusing the most prominent depth cues. In the context of this paper, only the measurement of linear perspective is studied since it is considered as one of the most prominent monocular depth cues.

The paper is structured as follows. Section 2 provides a concise review of depth perception in general, and the most important depth cues. Section 3 describes two approaches which have been developed to objectively measure the influence of the linear perspective on the perceived depth. In section 4 both approaches are evaluated on a publicly available 2D image dataset and their limitations are analyzed. Finally, section, 5 summarizes the work and discusses future research directions.

2. DEPTH PERCEPTION

There are many factors which contribute to the general understanding of the organization in depth of the different objects composing a scene (the depth layout). These can be decomposed into two distinct classes: the monocular depth cues and the binocular depth cues [1]. The *monocular depth cues* provide information on the depth using only a single view. They can be decomposed into two distinct classes: the static and motion-based cues. An illustration of different monocular depth cues is depicted in Figure 1. In addition to the monocular depth cues, the binocular vision provides *binocular depth cues*. The pupils of the two human eyes are shifted by

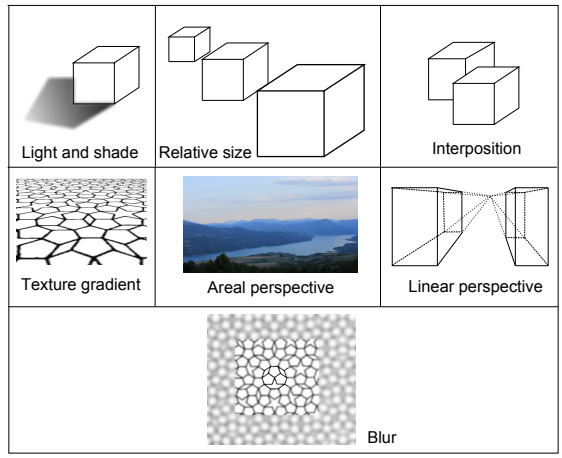

Fig. 1. Different types of monocular depth cues.

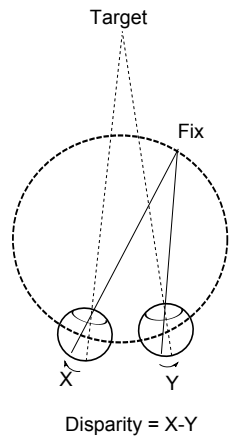

Fig. 2. Binocular disparity used for stereopsis.

approximatively 6.5 cm, which causes each retinal image to provide a slightly different view of the same scene. This difference between the two 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.

The question of how all these depth cues contribute to the general depth perception has been studied by Cutting & Vishton [6]. Their analysis included 9 distinct depth cues including occlusion, relative size, relative density, height in the visual field, perspective, motion perspective, binocular disparities, convergence and accomodation. Based on a comparison of their ordinal depth-threshold func-

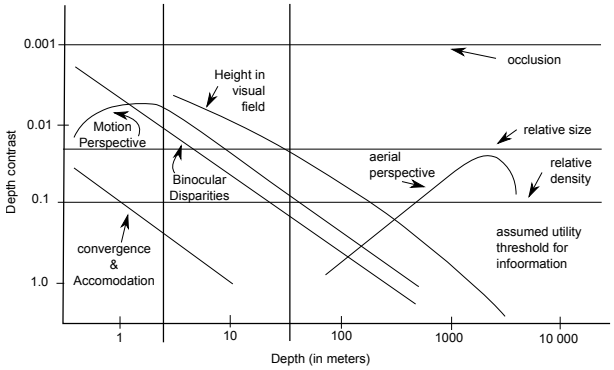

Fig. 3. Depth contrast perception as a function of the visualization distance [6].

tions they have shown that the influence of each depth cue on the overall depth perception depends largely on the distance (see figure 3). According to these variations they have partitioned the space around the observer into 3 concentric circles where specific depth cues are dominant. In the *personal space* (0–2 m) the most important depth cues are occlusions, retinal disparity, relative size, convergence and accommodation. In the *action space* (2-30 m) the depth cues can be ranked according to their importance in the following way: occlusion, height in the visual field, binocular disparity, motion perspective, and relative size. In the vista space (beyond 30 m) the only effective sources of information are the so-called pictorial cues including occlusion, height in the visual space, relative size and aerial perspective. While the derived rankings provide a way to understand the importance of the individial depth cues for certain depth ranges, a computational model for their integration is still missing and requires further studies.

3. LINEAR PERSPECTIVE

Linear perspective refers to the expansion of a 3D space [4] as it is illustrated in figure 4. The convergence of parallel lines such as the borders of the street in a visible vanishing point provides a strong perspective depth cue. On the other hand, the lack of vanishing lines for example in a row of trees or a vanishing point perpendicular to the camera view such as in a frontal view of a house provide very weak perspective depth cues.

In order to automatically predict the strength of linear perspective for a given image, local and

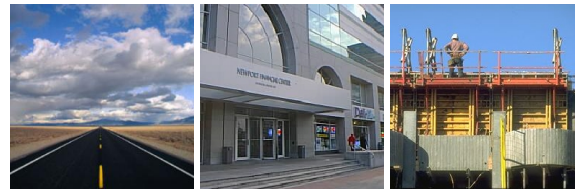

Fig. 4. From strong to weak linear perspective.

global image properties can be analyzed. One way is to detect and group visible lines and analyze the corresponding vanishing points. Another way is to ignore these local properties and characterize the scene layout globally by analyzing the statistical regularities of features within the image. Both approaches have been considered here, and the specific methods are described below.

3.1. Global layout properties (GLP)

The *global approach* is based on the analysis of global texture features and adopts the method proposed by Ross & Oliva [4] for scene description. A 2D image is divided into a 2×2 grid of non-overlapping rectangular regions which are described through a set of GIST features [7]. These features are obtained by convolving each image region with a set of Gabor-like filters and averaging the complex magnitudes at 4 scales and 8 orientations as it is illustrated for several examples in figure 5. Principal component analysis (PCA) is applied to reduce the dimensionality of the features from 128 to 24. Given a set of training images and the corresponding features a cluster weighted model (CVM) is trained to predict the strength of the perspective depth cues x within a test image. A CVM is essentially a generalization of a Gaussian mixture model (GMM) to linear regression problems. It consists of several clusters and their associated linear regression functions, which are combined to derive the predicted perspective according to the mixture proportions for a given sample.

3.2. Vanishing point model (VPM)

The *geometric approach* is based on the analysis of vanishing lines and their corresponding vanishing points, and adopts the unified model for geometric parsing by Barinova et al. [8]. It models the scene as a composition of geometric primitives at different levels (edges, lines, vanishing points) as

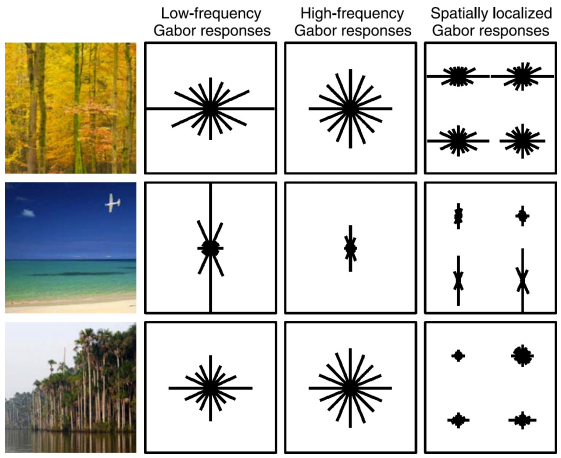

Fig. 5. Illustration of the global layout properties (GLP) with input image (left), global (middle) and local Gabor responses (right).

it is illustrated in figure 6. Line segments are detected by applying the line segment detector (LSD) by von Gioi et al. [9] on an image pyramid of 3 different scales and grouped into lines using the probabilistic Hough transform. The candidate vanishing point are derived using the J-linkage algorithm. In order to compute the final set of vanishing points along with the horizon and the zenith, the relationships between the different levels are explicitly recovered through a joint optimization process that does not require the Manhattan world assumption with 3 orthogonal directions to be fulfilled. By analysing the position of the detected vanishing points for a set of images a heuristic rule has been derived to predict the strength of the perspective cues. It is based on the observation that the perspective cues are strongest if one of the vanishing points is close to the image center and decreases if the vanishing points lie outside the image. Therefore, the strength of the perspective depth cues s is computed as $x = \frac{1}{d+1}$ given the distance d of the closest vanishing point to the image center.

4. EXPERIMENTAL RESULTS

4.1. Dataset

For the development and the evaluation of the two methods, the 2D image database created by Ross & Oliva [4] was used. It contains 7138 unique images from urban and natural outdoor environments. A few representative samples are shown in figure 7.

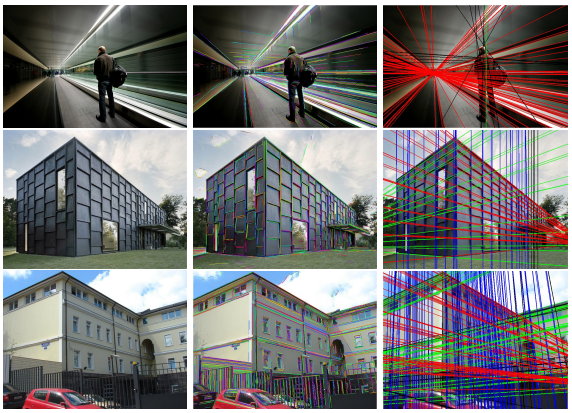

Fig. 6. Illustration of the vanishing point model (VPM) with input image (left), line segments (middle) and vanishing lines (right).

Fig. 7. Samples from the used 2D image dataset with urban (left) and natural environments (right).

For each of the images, 1 out of 14 observer has been asked to rate the degree of perspective, openness and depth on a scale from 1 to 6. A small subset of 838 images has been rated by a second observer, which can be used to consistency of the ratings. For the experiments the subjective scores of the perspective rating have been rescaled to the interval $[0, 1]$. They serve as ground truth y for the evaluation of the two methods.

4.2. Performance

The performance of the developed methods was evaluated by comparing the predicted objective values x to the ground truth values y obtained from the subjective test. Figures 8 and 9 provide scatter plots (subjective vs. objective scores) comparing the two methods for urban and natural environments, respectively. While the correlation between subjective and objective scores seems to be

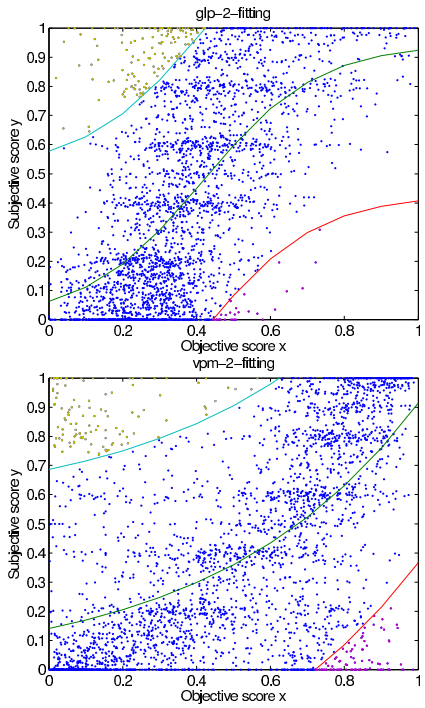

Fig. 8. Scatter plot (subjective vs. objective) of GLP (top) and VPM (bottom) for urban environments.

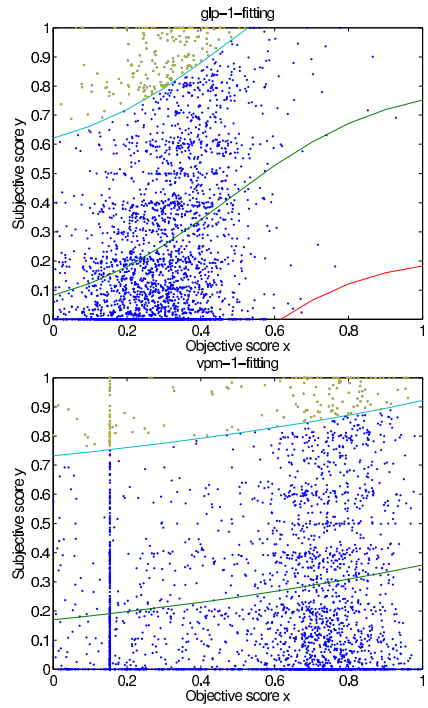

Fig. 9. Scatter plot (subjective vs. objective) of GLP (top) and VPM (bottom) for natural environments.

quite high for urban environments, it is very small for natural environments. This is confirmed by the computed Pearson correlation coeffcients after fitting a logistic function (green curve) [10]. For urban scenes the results are quite promising with 0.64 for the global approach (GLP) and 0.59 for the geometric approach (VPM). For natural scenes the performance drops considerably with correlation coefficients of 0.33 and 0.17, respectively.

4.3. Analysis

To better understand the limits of both approaches, the upper (cyan curve) and lower (red curve) prediction bounds, with a confidence level of 0.95, have been computed for the logistic function. The corresponding upper (yellow points) and lower (magenta points) outliers have been sorted according to their distance between the objective and subjective scores and have been visually analyzed. The figures 10 and 11 show a selection of the upper ($y > x$) and lower ($y < x$) outliers, respectively.

Analysing the upper outliers in figure 10 shows that the GLP method usually underestimates the strength of the linear perspective for scenes which

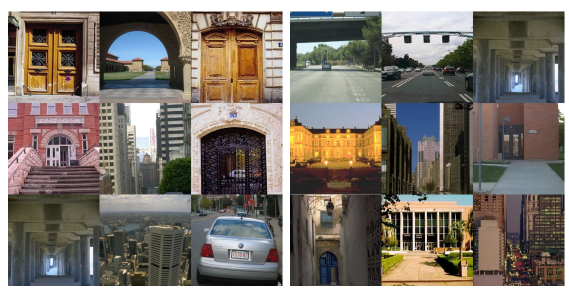

Fig. 10. Selected upper outliers of GLP (left) and VPM (right) for the urban environment.

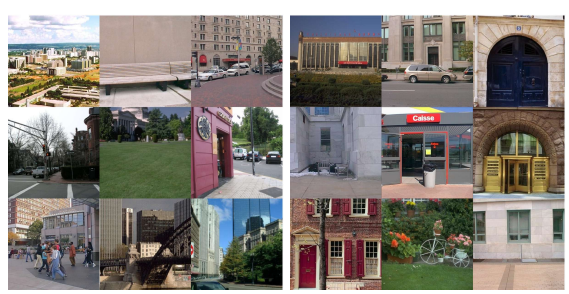

Fig. 11. Selected lower outliers of GLP (left) and VPM (right) for the urban environment.

contain strong textures throughout the whole image. On the other hand, the VPM method typically underestimates the strength of the linear perspective for scenes which contain only a few vanishing lines or which contain both parallel and perpendicular vanishing lines. For the lower outliers in figure 11 the GLP method seems to overestimate the strength of the linear perspective for scenes which contain both textured and homogeneous areas. The majority of the overestimates by the VPM method are caused by the non-parallel lines which are accidently considered as vanishing lines. Furthermore, both figures reveal an issue of the used dataset. Some large differences between the subjective and objective scores are actually caused by questionable subjective scores. An analysis of the images which have been rated twice shows that for 12% of the images the subjective ratings are inconsistent with a difference of 2 on a scale from 1 to 6. The performance can be expected to be higher if these ambiguous scores are excluded from the evaluation.

5. CONCLUSION

After providing a concise review of the most important aspects of depth perception, this paper explores global and geometric approaches for measuring linear perspective as one of the most important monocular depth cues. The global approach (GLP) uses texture features and a trained regression model to predict the strength of the perspective depth cues. The geometric approach (VPM) relies on the detection of vanishing lines and their corresponding vanishing points and uses a heuristic rule to relate their position to the linear perspective. While both approaches perform comparably well for urban environments, the performance drops considerably for natural environments. Especially, the geometric approach suffers from the lack of parallel lines in natural scenes. A detailed analysis of the outliers has revealed the complementary weaknesses of both approaches and motivates their integration as one future research direction. Another direction is the integration of additional monocular depth cues, which are assumed to be especially important when it comes to natural scenes. During the analysis questionable subjective scores in the used database have been observed which affect the measured performance. The lack of an alternative dataset shows that more comprehensive datasets with a large variety of monocular and binocular depth cues and reliable subjective ratings are needed.

References

[1] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, "Depth cues in human visual perception and their realization in 3D displays," in *Three-Dimensional Imaging, Visualization, and Display*, 2010.

[2] E Delage, H Lee, and A. Y Ng, "Automatic single-image 3d reconstructions of indoor manhattan world scenes," in *International Symposium of Robotics Research*, 2005.

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

[4] M.G. Ross and A. Oliva, "Estimating perception of scene layout properties from global image features," *Journal of Vision*, vol. 10, no. 1, 2010.

[5] ITU-T, "Audiovisual quality in multimedia services: Subjective video quality assessment methods for multimedia applications," Tech. Rep. P.910, 1999.

[6] J.E. Cutting and P.M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth," *Perception of space and motion*, vol. 5, pp. 69–117, 1995.

[7] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, 2001.

[8] O Barinova, V. Lempitsky, E. Tretyak, and P. Kohli, "Geometric image parsing in man-made environments," in *ECCV*, 2010.

[9] R.G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 4, pp. 722 –732, april 2010.

[10] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep. BT.500-11, ITU-R, 2002.