Mikhail Erofeev, Dmitriy Vatolin, Alexander Voronov, Alexey Fedorov,
"Toward an Objective Stereo-Video Quality Metric: Depth Perception of Textured Areas,"
International Conference on 3D Imaging (IC3D),
pp. 1–6, 2012. doi:10.1109/IC3D.2012.6615120

TOWARD AN OBJECTIVE STEREO-VIDEO QUALITY METRIC: DEPTH PERCEPTION OF TEXTURED AREAS

Mikhail Erofeev, Dmitriy Vatolin, Alexander Voronov, Alexey Fedorov

Department of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia

ABSTRACT

Objective stereo quality metrics are required by scientific and manufacturing communities to properly capture, process, encode and transmit stereo video. Over recent years several stereo-video quality metrics have been proposed. In this paper we address the problem of constructing an objective method for detecting image areas sensitive to stereo distortions. We describe a subjective testing methodology whose goal is to determine how image texture affects human depth perception. We then construct our pixelwise subjective method of scoring the sensitivity of image areas to depth-map distortions. Finally, we perform subjective validation of the results. Possible applications of the proposed method include incorporating the resulting predictions into existing stereo quality metrics and encoding algorithms to achieve better correlation with subjective measurements.

Index Terms— Stereo vision, stereo image processing, image analysis

1. INTRODUCTION

Today, content in 3D format is becoming increasingly popular. To avoid a low-quality end-user experience, the video industry requires methods for objectively measuring 3D-content quality. In this case subjective quality-measurement methods cannot be applied because of their high cost and time inefficiency. Over recent years several approaches for evaluating stereo-video quality have been proposed by different authors [4, 2, 5, 1, 8]. We discuss these approaches in more detail in the next section.

In this paper we propose an objective way to determine the sensitivity of different image areas to depth-map distortion. Our method is based on data collected during several subjective tests whose purpose was to determine how different textures affect human depth perception. This testing illustrated dramatic variations in depth perception for different image textures. Below, we describe the testing procedure and discuss the results. The differences in depth perception mean that texture characteristics should be taken into account when constructing a stereo quality metric for better correlation with subjective measurements. Thus, in this research we avoid addressing the entire problem of developing a stereo-image quality metric; instead we focus on the significance of depth-map artifacts for different image areas. Our results, however, can be merged into new quality metrics, and we expect them to be especially useful for metrics that are based on a single image plus a depth-data representation.

Our discussion begins with related work and then moves on to our subjective testing methodology, the process of constructing the texture scoring algorithm and the algorithm's results.

2. RELATED WORK

Several efforts to directly apply existing 2D quality metrics to 3D quality measurement have been undertaken. The authors of [4] evaluated PSNR, VQM and SSIM quality metrics and reached the conclusion that VQM has a good correlation with subjective measurements, but subjective results should remain a "gold standard." In [2], the authors performed subjective evaluation of the SSIM, UQI, C4, and RRIQA quality metrics and tried different ways of combining their scores for the left and right images. Their conclusion was that despite good correlation of the subjective results with the 2D-image quality metrics, extensive additional work addressing binocular distortions is necessary.

In [5], the authors considered binocular distortions using 3D-DCT structure, which consists of similar blocks from the left and right views. Another full-reference metric proposed in [1] takes monoscopic and binocular artifacts into account separately. Monoscopic distortions were computed using a cyclopean image, and a disparity map was used to compute binocular artifacts. Both approaches very closely match current theories about how the human visual system (HVS) operates.

For the present work we propose a way to model how the HVS deals with different textures. This information can be useful in constructing new and more-precise models of the HVS.

A no-reference metric, proposed in [8] by Anish Mittal et

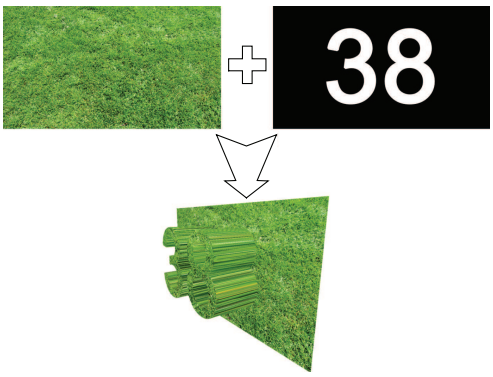


Fig. 1: Scheme of test-pattern construction.

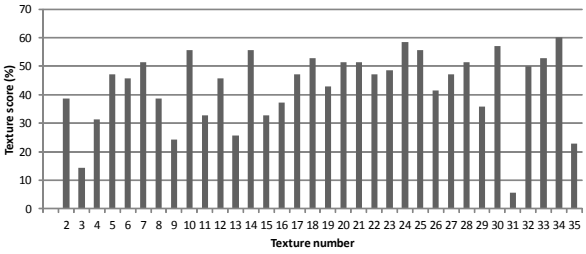


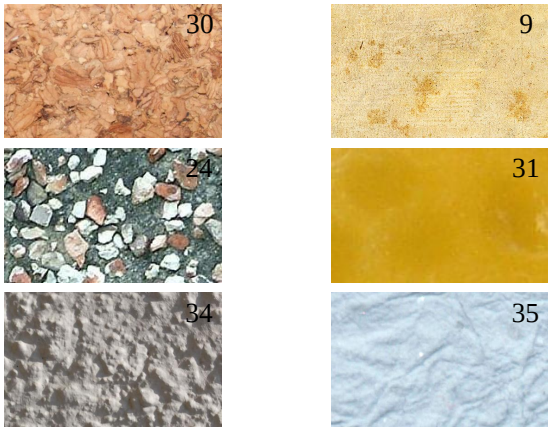
Fig. 2: Subjective testing results.

al., was constructed using machine-learning algorithms. The authors showed good correlation of their metric with [3] and other subjective data sets. Our work uses similar ideas to fit subjective data.

3. SUBJECTIVE TESTING

To determine how different textures affect human depth perception, we collected a set of 35 textures with different characteristics. The resolution of each texture was 1920×1080. These textures were then used to generate test patterns similar to the random-dot stereo images described in [6]; the only difference was that we used a texture instead of random dots. In each test pattern we encoded two digits using three-pixel shifts. The size of two digit bounding box was 400×300 and it was located randomly on the entire texture. Humans can recognize numbers only when such patterns are displayed as stereo images; they fail to recognize these numbers when only one view is shown. Figure 1 shows pattern-construction scheme.

A total of 90 people were invited to take part in the subjective perception testing. The average age of the participants was 18.4 years. All participants were asked to provide information about eye diseases they had and optical power of their glasses and use them, if glasses were prescribed by doctor. We used a cinema-like projector system with polarization filters to display the test patterns. We showed the first test image to participants and indicated the correct response, thus offering a better explanation of the testing procedure. All other images were shown sequentially for 20 seconds each. Respondents were asked to write down the numbers that they could see. Finally, we collected all the written response forms and computed the percentage of correct answers for each texture (which we refer to as the “texture score”). Originally we were going to remove all outlying results from the collected data but there were no sufficient outliers, even the participant with color blindness (according to information) had typical not outlying result.



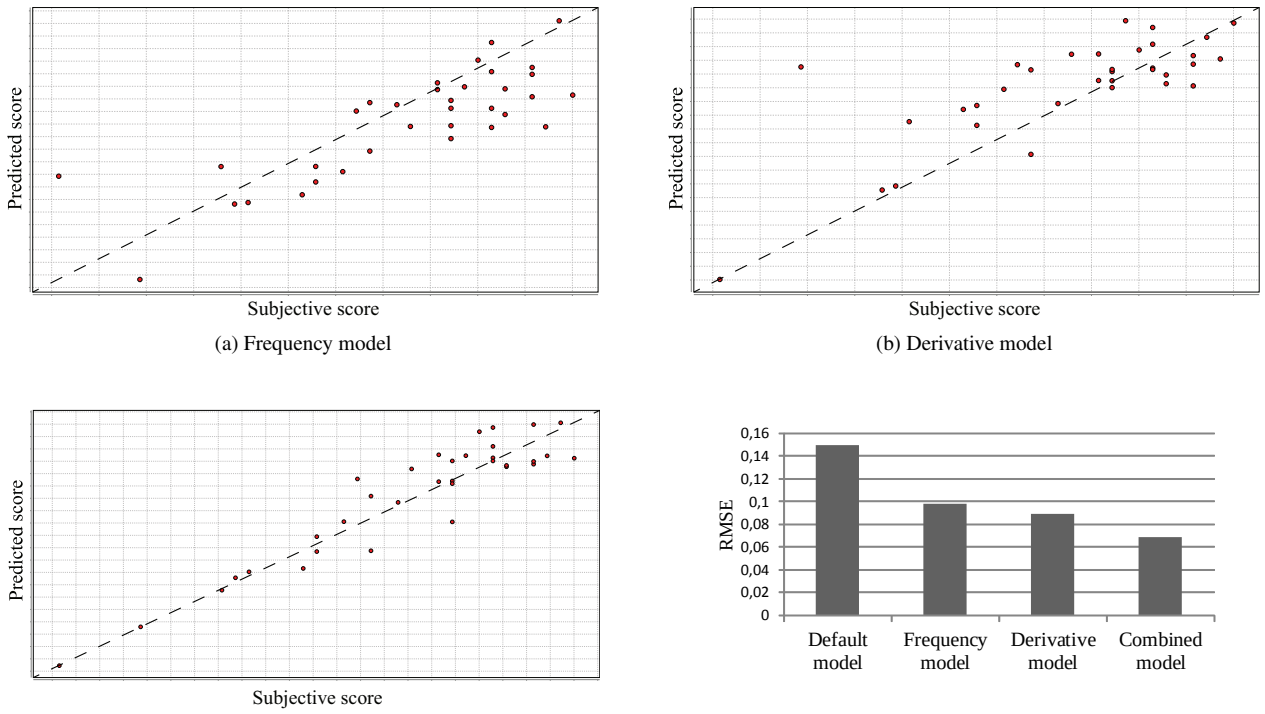
(a) Highest score (b) Lowest score

Fig. 3: Examples of textures with the highest (a) and lowest (b) subjective scores. Numbers on textures correspond to numbers in Fig. 2

Before developing the objective method, we performed some manual analysis of the collected data. Figure 2 shows the distribution of scores for different textures. A noticeable feature is that the score depends highly on texture. Figure 3 shows examples of textures with the lowest and the highest scores. Obviously textures with a low score have no clearly visible structure or edges; on the other hand, textures with a high score have clearly visible structure and edges.

4. OBJECTIVE METHOD CONSTRUCTION

At this point, the main goal of our research was to develop a method that estimates a value close to the texture score for the given texture image–in other words, a method that gauges texture sensitivity to depth-map distortions. To accomplish this task, we decided to use machine-learning algorithms, then solved the two-fold problem of selecting a good feature set to describe textures in the data set and of selecting a suitable method for solving the regression problem.



(a) Frequency model

(b) Derivative model

(c) Combined model

(d) RMSE comparison

Fig. 4: Curve-fitting results. Parts (a)–(c) show how different models fit the subjective data set. X axis indicates subjective score – texture score (percentage of correct answers for each texture) measured during subjective testing. Y axis indicates predicted score – output of the obtained linear function based on the appropriate feature set. Part (d) shows the RMSE of the cross-validation check for each model.

4.1. Feature-set selection

Owing to the small learning data set (34 textures, one texture was withdrawn because it was used for better illustration of the subjective testing procedure), we were forced to use a small feature set to avoid overfitting. We used two groups of features: frequency features and derivative features. Both of these groups are designed to take into account a range of frequencies from different bands. Furthermore, we consider all images to be monochromatic, having pixel intensities between 0 and 1.

To compute the frequency features, an entire texture image was smoothed using a Gaussian kernel ($\sigma = 0.5$) to remove high-level noise. Then for each image pixel ($I(x, y)$) and its neighbors in horizontal direction, eight features ($f_i^{freq}(x, y)$) were computed using formula (1).

$$f_i^{freq}(x,y) = \frac{1}{|l_i|} \sum_{j \in l_i} |D_j\left[I(x,y)\right]| \tag{1}$$

Here, D_j is the jth 64-point discrete Fourier transform (DFT) coefficient. The DFT was computed for each pixel and its horizontal neighborhood. Also, $\{l_i\}$ divides the first 32 coefficients of the DFT into eight segments with equal length. Thus, each f_i^{freq} corresponds to an average absolute amplitude of frequencies in the ith band. Finally, to get frequency features for the entire texture we use the following expression:

$$F_i^{freq} = \frac{1}{|I|} \sum_{(x,y) \in I} f_i^{freq}(x,y). \tag{2}$$

Derivative features are based on the image's second horizontal derivatives and are computed using formula (3).

$$d_i(x,y) = \frac{I(x-i,y) + I(x+i,y)}{2} - I(x,y)$$
$$f_i^{deriv} = \min(|d_i(x,y)|, \theta) \tag{3}$$

Therefore, f_i^{deriv} corresponds to the absolute value of the second derivative computed in the horizontal direction for a pixel neighborhood width of $2i$, and it is limited by a threshold ($\theta = 0.009$). We limit derivative values using a threshold to avoid overrating images with very hard edges, since our assumption is that the HVS perceives no difference in edge hardness if the edge is already visible. Finally, to get feature values for the whole texture, we also compute the average of pixelwise features using formula (4).

$$F_i^{deriv} = \frac{1}{\|I\|} \sum_{(x,y) \in I} f_i^{deriv}(x,y) \tag{4}$$

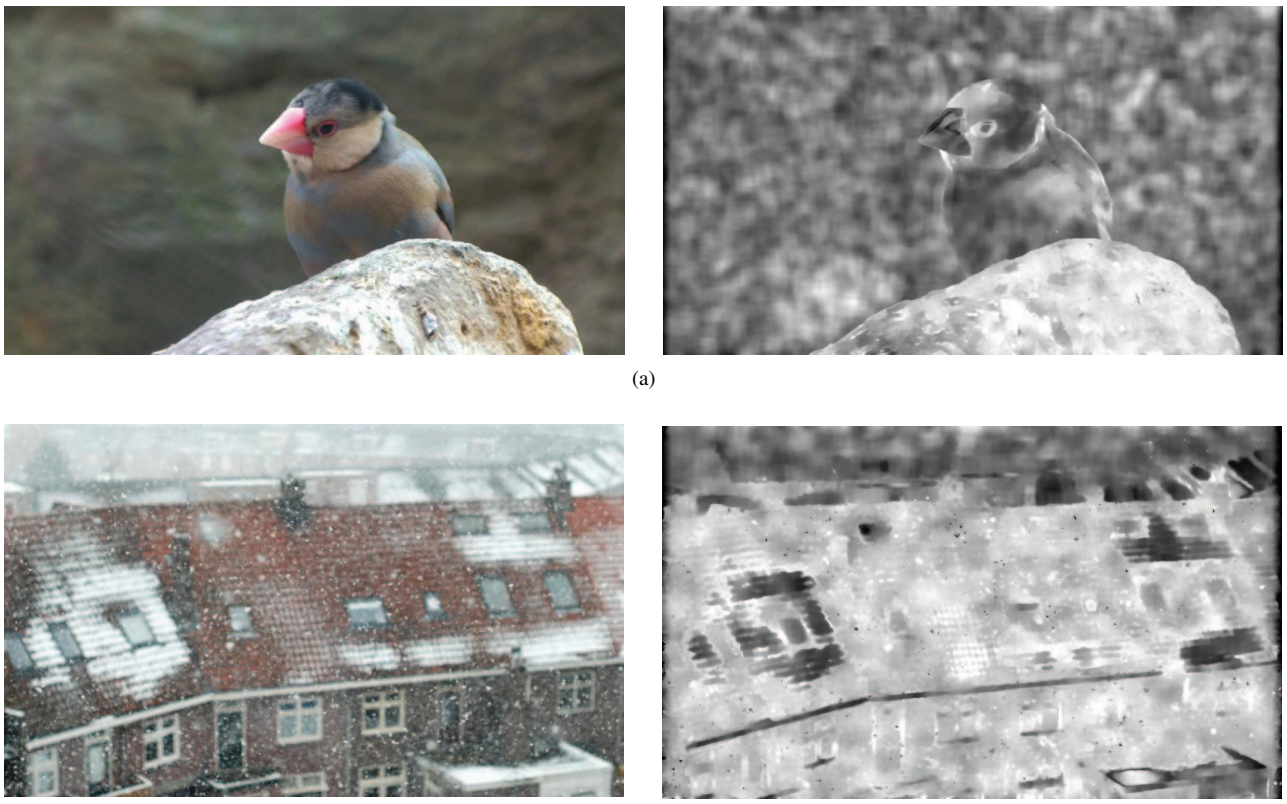


Fig. 5: Method application to natural images. Source images are depicted on the left side and results on the right side.

Selection of both feature sets is motivated by the idea to estimate granularity of the textures and amount of granules of different sizes. In different ophthalmologic researches were shown that neurons in human brain are very sensitive to edges in case of stereoscopic perception, also it was shown [9] that human ability to correctly perceive stereoscopic stimulus is highly correlated with its spatial frequency. Thereby features based on measurement of texture granularity and granules size have some medical justification.

4.2. Regression problem solution

At this point of the research our goal was to obtain function that for the given texture computed previously would output value close to texture score measured during the subjective testing. Considering that our data set was small, we avoided complex models and used linear regression to fit the subjective data collected previously. We also applied machine-learning algorithms, like neural networks and curvilinear regression, but their performance was significantly lower. Root-mean-squared error (RMSE) was applied in the cross-validation check to control the quality of the final model. For better comparison, we also tested a model that does not take into account any features but that approximates the training set as an average score value (default model).

For the frequency, derivative and combined (union of frequency and derivative features) feature sets, three separate linear models were constructed and tested. The RMSE values for the default, frequency, derivative and combined models in the cross-validation check were 0.15, 0.098, 0.089 and 0.069, respectively. Figure 4 (a)–(c) illustrate how these models fit the subjective data, and Figure 4 (d) compares each model's RMSE with that of the default model. The model relying on the combined features set had the lowest RMSE, so it was used to construct the final pixelwise scoring method, which is described below.

4.3. Final pixelwise scoring

The final scoring method works as follows:

1. Blur entire image with Gaussian kernel ($\sigma = 0.5$)
2. For each pixel:
   (a) Compute frequency features
   (b) Compute second-derivative features
   (c) Apply combined linear model described in Subsection 4.2

3. Perform cross-bilateral filtering (described in [7]) on the map of score values for each pixel using the source image

The main goal of the first step is removal of high-frequency noise, which distorts DFT coefficients and derivative values. The second step involves separate computation of the score for each pixel. The usage of the bilateral filtering in the last step is motivated by the spatial inconsistency of the metric values (especially along the edges) and its goal is making the final result smoother preserving strong edges on the entire image. The results of applying the scoring method to different images are depicted in Figure 5 and are discussed in the next section.

5. METHOD ANALYSIS AND VALIDATION

Figure 5 shows the results of the proposed method (based on combined features set). In the resulting images, brighter points should be interpreted as a higher score–in other words, these points are very sensitive to stereo distortions.

Figure 5 (a) shows that the stone with a very strong texture has a high score, whereas the blurred background, which lacks details, has a low score. These results are very close to the intuitive understanding of human stereo perception.

Furthermore, to validate our results we performed one more subjective test. The depth maps of a set of test images were distorted in the regions with the highest and lowest scores (Figure 6 shows examples of distorted depth maps and their corresponding image). To distort depth maps we put small rectangular on them with different depth value in the random position. Such selection of distortion type was motivated by the goal to make it as noticeable for human as possible, thus observer’s inability to find such artifact would mean that area where distortion is located is highly insensitive to depth map distortions. Observers were then asked to locate these distortions. The results of this experiment were mostly in agreement with the scores predicted by the method. Four participants were shown eight distorted images, for seven of the images, the participants were unable to find distortions in regions having low scores, whereas they easily found artifacts in areas with high scores. For one image the proposed method predicted that distortion would be clearly visible, but none of the participants were able to see it. This incorrect prediction can be explained by the lack of dark textures in our training data set and by the absence of color features in the feature list; thus, the curve-fitting algorithm lacked sufficient data to deal correctly with this case.

(a) Source image

(b) Low-score distortion

(c) High-score distortion

Fig. 6: Example of source image (a) and distorted depth maps (b)–(c).

6. DIRECTIONS FOR FURTHER WORK

Despite the good correlation of our method’s predictions with experimental data, as well as partial validation of its results, extensive further work should be pursued. A more representative set of textures should be collected and scored using the subjective scoring procedure described in Section 3. Such a data set will allow addition of more items to the feature set and will enable use of more-complex models.

To determine how an image’s color and brightness affect human stereo perception, another subjective test should be conducted. During this test, participants should be shown a given texture with different global color transformations. For example, these transformations might include a global brightness decrease or increase, transformation of a hue component or reduction in the amount of red color.

Also, the current method is highly dependent on the display device and viewing distance because the only unit of measurement was the pixel, so all computations should be converted to device-independent measurement units.

Finally, practical applications of the proposed method should be studied, such as construction of a full-reference quality metric based on a 2D plus depth-map data representation or involving our method in existing stereo quality metrics.

7. CONCLUSION

In this paper we addressed the problem of constructing an objective method to detect image areas sensitive to stereo dis-

tortions. Our research involved performing several subjective experiments, the results of which were used to construct the final method by way of machine-learning algorithms. The proposed method was then validated using one more subjective test.

8. ACKNOWLEDGEMENTS

This work is partially supported by the Intel/Cisco Video-Aware Wireless Network (VAWN) Program and by grant 10-01-00697a from the Russian Foundation of Basic Research.

9. REFERENCES

[1] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G.B. Akar. Towards compound stereo-video quality metric: a specific encoder-based framework. In *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on*, pages 218 –222, 2006.

[2] Patrizio Campisi, Patrick Le Callet, and Enrico Marini. Stereoscopic images quality assessment. In *15th European Signal Processing Conference, 2007 EURASIP*, pages 2110 –2114, Sep. 2007.

[3] Lutz Goldmann, Francesca De Simone, and Touradj Ebrahimi. Impact of acquisition distortions on the quality of stereoscopic images. In *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2010.

[4] C.T.E.R. Hewage, S.T. Worrall, S. Dogan, and A.M. Kondoz. Prediction of stereoscopic video quality using objective quality models of 2-d video. *Electronics Letters*, 44(16):963 –965, Jul. 31 2008.

[5] Lina Jin, Atanas Gotchev, Atanas Boev, and Karen Egiazarian. Validation of a new full reference metric for quality assessment of mobile 3dtv content. In *19th European Signal Processing Conference, 2011 EURASIP*, pages 1894 –1898, Sep. 2011.

[6] B. Julesz. Binocular depth perception without familiarity cues. *Science*, 145:35 –362, 1964.

[7] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007)*, 26(3), 2007.

[8] A. Mittal, A.K. Moorthy, J. Ghosh, and A.C. Bovik. Algorithmic assessment of 3d quality of experience for images and videos. In *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*, pages 338 –343, Jan. 2011.

[9] Clifton M. Schor and Ivan Wood. Disparity range for local stereopsis as a function of luminance spatial frequency. *Vision Research*, 23(12):1649 – 1654, 1983.