A perceptual quality metric for high-definition stereoscopic 3D video

F. Battisti[a], M. Carli[a], A. Stramacci[a], A. Boev[b], A. Gotchev[b]

[a]Department of Engineering, Universita' degli Studi Roma TRE, Roma, Italy;

[b]Tampere University of Technology, Tampere, Finland.

ABSTRACT

The use of 3D video is growing in several fields such as entertainment, military simulations, medical applications. However, the process of recording, transmitting, and processing 3D video is prone to errors thus producing artifacts that may affect the perceived quality. Nowadays a challenging task is the definition of a new metric able to predict the perceived quality with low computational complexity in order to be used in real-time applications. The research in this field is very active due to the complexity of the analysis of the influence of stereoscopic cues. In this paper we present a novel stereoscopic metric based on the combination of relevant features able to predict the subjective quality rating in a more accurate way.

Keywords: Quality Metric, Stereoscopic 3D Video, Video Transmission

1. INTRODUCTION

3D video is a growing technology that can potentially affect many fields, such as entertainment, military simulations, movies, medical applications and 3DTV[1–4] . The success of 3D imaging world relies on the ability of 3D systems to provide an added-value compared to conventional monoscopic imaging (i.e. depth feeling or parallax motion) coupled with high image quality contents. However, the 3D data may be affected by errors that can be originated in several steps of the communication chain, from its generation, to processing, transmission, and rendering. Those errors lead to artifacts that may be quite different, both for realization and impact, from the ones affecting 2D videos to which we are accustomed. An analysis of most common 3D artifacts can be found in[5] . Quality assessment tools are thus needed for ensuring reliable quality evaluations.
Despite the advances in view modeling and synthesis, less efforts have been devoted to develop algorithms for assessing the visual quality of a stereoscopic 3D video. One of the most challenging tasks, is the design of a metric able to predict as close as possible the perceived quality and, at the same time, to require a low computational cost to be adopted in real-time applications. Quality assessment can be performed by means of subjective tests; however, they are expensive, time consuming, and the collected results may be affected by factors which can significantly infer their reliability, such as loss of concentration of users during the test. To reduce the above-mentioned drawbacks, objective metrics able to mimic the human judgement are being developed.

Several studies have been performed for evaluating the quality of 2D videos and images whereas for 3D videos there are still many difficulties due to the more complex influence of the stereoscopic cues. Efforts have been devoted for evaluating the effectiveness of 2D quality metrics when applied to stereoscopic data: in these simple approaches, a 2D metric is applied to each channel of the stereo video and then the overall 3D video quality is obtained by averaging the separate scores[6,7] . The collected results show that these models do not resemble the binocular mechanisms of the human visual system resulting in low correlation with the subjective scores. To improve the metric effectiveness, typical 3D factors have been included in novel metrics[8–11] . Studies have also been carried out for creating stereoscopic video databases containing scenes with heterogeneous content and different capture parameters as in[12,13] .

In this work the goal is to define a new metric able to accurately predict the subjective judgement when applied to high resolution 3D stereoscopic videos. The proposed scheme is inspired by[14] , where the authors proposed an effective 3D quality assessment metric for the mobile device scenario. Here, we select relevant features with the aim of tuning the quality metric to a high-definition stereoscopic video scenario. In this work a model for assessing stereoscopic image quality is defined by considering: the quality of the cyclopean view, the presence of binocular rivalry, and the presence of binocular depth. Basically, the quality of the single (cyclopean) image obtained by merging the left and right views, the influence of binocular rivalry on visual comfort, and the impact of the depth on correct perception of the 3D scene geometry are considered. To quantify these components, features extracted from the data are analyzed and combined. Here, we select the features whose combination better matches the subjective scores. Since existing 2D quality metrics are

based on the computation and weighting of low level features, for computing our features we exploit existing state of the art 2D quality metrics.
The rest of the paper is organized as follows. In Section 2, the details of the proposed method, with the adopted 3D factors considered are presented and the selected features, adopted in different 2D quality metrics, are briefly described. In Section 3 the steps performed for defining the overall metric are described and the achieved results are discussed. Finally, in Section 4 the conclusions are drawn.

2. PROPOSED METHOD

As previously mentioned, the proposed approach relies on the combination of features extracted by exploiting 2D quality metrics in order to define a reliable 3D metric able to predict in an accurate way the MOS. In order to achieve this goal, the quality of the cyclopean view, the binocular rivalry, and the binocular depth have been considered. In the following, the details about the exploited vision models and the selected quality metrics are reported.

2.1 Vision models

Let us define with I_{ref} and I_{dis} respectively the original and distorted views; we can thus define:

- Cyclopean view (CV): it is given by the overlapping between left and right view. Since we want to evaluate its quality, first it is computed as in[14] , and then its quality is evaluated by means of three models, all based on the use of any Quality Assessment metric (QA) able to measure the similarity among images:
  - $CV_1 = QA(I^{cyc}_{ref}, I^{cyc}_{dis})$;
  - $CV_2 = \frac{\sum_{i=1}^{N_{blk}} MAX(q_i^L, q_i^R)}{N_{blk}}$ where $q_i^L = QA(A_{ref}, A_{dis})$ and $q_i^R = QA(B_{ref}, B_{dis})$;
  - $CV_3 = \frac{\sum_{i=1}^{N_{blk}} (q_i^L + q_i^R)/2}{N_{blk}}$ where $q_i^L = QA(A_{ref}, A_{dis})$ and $q_i^R = QA(B_{ref}, B_{dis})$;

  where I^{cyc}_{ref} and I^{cyc}_{dis} are the original and distorted cyclopean views, A and B are $k \times k$ blocks extracted from the left and right views, and N_{blk} is the total number of blocks in which the images are partitioned. The B block of the right view is selected taking into account the shift between the views using the so called *block grouping* procedure[14] .

- Binocular rivalry (BR): it occurs when the eyes try to focus on a single point in a scene as a result of two slightly different views. Even though occlusions are a natural source of artifacts, the major contribute to BR is given by the distorted views only. The reference view is not taken into account for this model. It can be computed as:

$$BR = \frac{\sum_{i=1}^{N_{blk}} QA(A_{dis} + B_{dis})}{N_{blk}} \tag{1}$$

- Binocular depth (DQ): it takes into account the amount of dep th in different stereoscopic videos. The binocular depth assessment is given by:

$$DQ = QA(\Delta_{ref}, \Delta_{dis}) \tag{2}$$

  where Δ_{ref} and Δ_{dis} are the disparity maps of the reference and distorted views.

2.2 Selected quality metrics

The following quality measures have been considered in the quality metric design:

- Mean Squared Error (MSE): it is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. MSE measures the average of the squares of the *errors*. It can be evaluated as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\widehat{I_{ref}} - I_{dis})^2;$$

- Sum of Squared Differences (SSD): it is a normalized version[15] of the MSE and it penalizes local intensity variations in textured areas:

$$SSD = \frac{1}{MN}\sum_i \sum_j \frac{(I_{dis} - \widehat{I_{ref}})^2}{|grad(I_{dis})|^2 - 1}$$

where $grad(u)$ is the gradient value of input signal, and M, N represent the size of the image;

- Peak Signal-to-Noise Ratio HVS (PSNR-HVS$_{c1}$)[16] : it is Peak Signal to Noise Ratio taking into account Contrast Sensitivity Function (CSF);
- Peak Signal-to-Noise Ratio HVS with Masking function (PSNR-HVS$_{c2}$)[17] : PSNR function taking into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions;
- Feature Similarity Index (FSIM$_{c1}$): it is based on the exploitation of physiological and psychophysical studies showing that visually discernible features coincide with those points where the Fourier waves at different frequencies have congruent phases[18] . That is, at points characterized by high Phase Congruency (PC), highly informative features can be extracted. Another feature considered in FSIM is the image Gradient Magnitude (GM), computed for encoding contrast information. PC and GM are complementary and they reflect different characteristics of the HVS in assessing the local quality of the input image. The similarity measure for $PC(I_{ref})$ and $PC(I_{dis})$ is defined as follows:

  $$S_{PC} = \frac{2PC(I_{ref})PC(I_{dis}) + T_1}{PC(I_{ref})^2 PC(I_{dis})^2 + T_1}$$

  where T_1 is a positive constant proportional to the dynamic range of PC values to ensure stability. The similarity measure for $GM(I_{ref})$ and $GM(I_{dis})$ is defined as follows:

  $$S_G = \frac{2GM(I_{ref})GM(I_{dis}) + T_2}{GM(I_{ref})^2 GM(I_{dis})^2 + T_2}$$

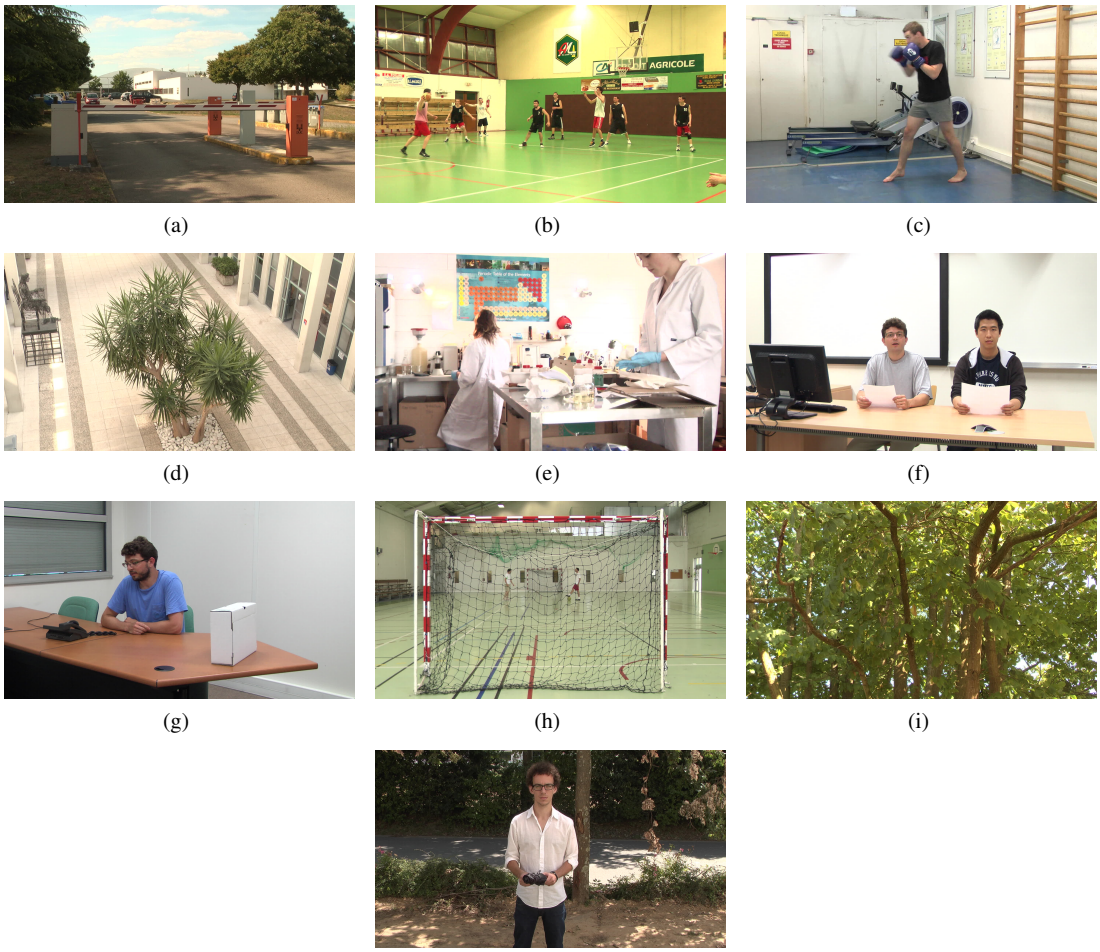
  where T_2 is a positive constant proportional to the dynamic range of GM values to ensure stability. In order to get the similarity between $f_1(x)$ and $f_2(x)$ the previous components are combined together:

  $$S_L(x) = [S_{PC}(x)]^\alpha [S_G(x)]^\beta$$

  where α and β are parameters used to adjust the relative importance of the two components. However, different locations have different contributions to HVS perception of the image. For example, edge locations convey more crucial visual information than the locations within a smooth area. Since human visual cortex is sensitive to phase congruent structures, the PC value at a location can reflect how likely it is a perceptibly significant structure point. Intuitively, for a given location (x, y), if any of $I_{ref}(x, y)$ and $I_{dis}(x, y)$ has a significant PC value, it implies that the pixel in position (x, y) will have a high impact on HVS in evaluating the similarity between I_{ref} and I_{dis}. Therefore, we use $PC_m(x, y) = max(PC_1(x, y), PC_2(x, y))$ to weight the importance of SL(x,y) in the overall similarity between I_{ref} and I_{dis}, and accordingly the FSIM index between I_{ref} and I_{dis} is defined as

  $$FSIM = \frac{\sum_{x,y=\omega} S_L(x,y) PC_m(x,y)}{\sum_{x,y=\omega} PC_m(x,y)}$$

  where ω is the spatial domain;
- Feature Similarity Index (FSIM$_{c2}$): FSIM computed by considering the PC component only;
- Feature Similarity Index (FSIM$_{c3}$): FSIM computed by considering the GM component only;
- Structural Similarity Index (SSIM)[19] : it considers image degradation as a perceived change in structural information. Structural information relies on the hypothesis that the pixels have strong inter-dependencies especially when they are spatially close. These dependencies carry important information about the structure of the objects in the visual scene. In short, it provides a quality evaluation based on tree different characteristics: *luminance*, *contrast* and *structure*. The first SSIM-based feature has been obtained by applying SSIM to the luminance component (Y) of the image;
- Structural Similarity Index (SSIM$_{c2}$): SSIM applied to the Cb color component;
- Structural Similarity Index (SSIM$_{c3}$): SSIM applied to the Cr color component.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 1: Left sample frames extracted from the videos in the DB.

3. METRIC DESIGN AND RESULTS

In order to design the proposed metric, the stereoscopic HD database *Nantes-Madrid-3D-Stereoscopic-V1 (NAMA3DS1)*[13] has been used. The database contains 10 original videos (SRC, SouRCe) that have been captured with a Panasonic AG-3DA1E twins-lens camera, with 60 mm distance between the lens. Twins lenses are adjusted to avoid vertical and angular rotations and brightness mismatching. All videos have Full HD 1080p resolution and frame rate of 25 fps. The stereoscopic sequences are chosen with different motion, environments (outdoor, indoor) and depth. Sample frames from the video sequences are in Figure 1. From each SRC, 9 PVSs (Processed Video Sequence) have been created by affecting the SRCs with 9 different artifacts (blocking, down-sampling, edge enhancement, and combinations of these artifacts). In order to provide a ground truth for quality metric assessment, subjective tests have been performed for collecting the Mean Opinion Score (MOS). 28 subjects took part to the subjective experiments and each participant has been asked to evaluate the perceived quality in a range from 1 to 5 where 1 corresponds to extremely poor quality and 5 to extremely good as summarized in Table 1.

3.1 Features computation

The selected QAs have been used in the vision models above mentioned and tested on the SRCs and PVCs to obtain the features named according to the scheme in Table 2 . The fitting of each feature f_j with the MOS is evaluated by computing

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 1: MOS scale example

	CV_1	CV_2	CV_3	BR	DQ
MSE	f_1	f_{11}	f_{21}	f_{31}	f_{41}
SSD	f_2	f_{12}	f_{22}	f_{32}	f_{42}
PSNR-HVS$_{c1}$	f_3	f_{13}	f_{23}	f_{33}	f_{43}
PSNR-HVS$_{c2}$	f_4	f_{14}	f_{24}	f_{34}	f_{44}
FSIM$_{c1}$	f_5	f_{15}	f_{25}	f_{35}	f_{45}
FSIM$_{c2}$	f_6	f_{16}	f_{26}	f_{36}	f_{46}
FSIM$_{c3}$	f_7	f_{17}	f_{27}	f_{37}	f_{47}
SSIM$_{c1}$	f_8	f_{18}	f_{28}	f_{38}	f_{48}
SSIM$_{c2}$	f_9	f_{19}	f_{29}	f_{39}	f_{49}
SSIM$_{c3}$	f_{10}	f_{20}	f_{30}	f_{40}	f_{50}

Table 2: Features denomination.

the Spearman correlation, as reported in Table 3.

As can be noticed, the best result is achieved for f_{25} for which the correlation value with the MOS is 0.81. In the last row of Table 3 the value of SROCC$_1$ corresponds to the correlation between the MOS and the combination of all features for each vision model. In this case it can be noticed that CV shows the best behavior with a maximum correlation value of 0.91. Since our aim is to exploit the characteristics of all models in order to have a more general approach, we decided to move forward towards the combination of the considered features.

3.2 Feature combination

The features combination is obtained through a *linear regression* between the selected features and the reference vector (the MOS). The goal is to find the parameters that are able to minimize the error between our selected features and the MOS. To the aim of selecting the minimum number of features needed to design our new metric, the *sequential feature correlation*[14] , has been applied. It can be summarized as follows:

	CV_1	CV_2	CV_3	BR	DQ
MSE	-0.33	-0.47	-0.50	0.10	-0.54
SSD	-0.47	-0.70	-0.72	-0.08	-0.57
PSNR-HVS$_{c1}$	0.38	0.14	0.16	-0.20	0.56
PSNR-HVS$_{c2}$	0.42	0.14	0.16	-0.20	0.57
FSIM$_{c1}$	0.54	0.78	0.82	-0.22	0.66
FSIM$_{c2}$	0.59	0.76	0.75	-0.16	0.65
FSIM$_{c3}$	0.43	0.57	0.56	-0.19	0.66
SSIM$_{c1}$	0.42	0.79	0.77	-0.11	0.66
SSIM$_{c2}$	0.47	0.50	0.50	0.11	0.65
SSIM$_{c2}$	0.42	0.79	0.77	-0.10	0.52
SROCC$_1$	**0.87**	**0.92**	**0.91**	**-0.85**	**0.83**

Table 3: Correlation values between MOS and the selected features.

1. The feature that exhibits the highest individual correlation is chosen;
2. a second feature is added;
3. the correlation between the first feature and all the others is computed;
4. the feature that shows the highest correlation is added to the set;
5. a third feature is added;
6. with an iteration, the third feature providing the highest correlation is found;
7. the procedure is iterated until no further improvement is obtained.

The outputs of the linear regression are reported in Figure 2. It can be noticed that 5 features are needed to reach a correlation value of 0.9. Based on the analysis of those results, 5 features have been chosen based on the following requirements: correlation larger than 0.9, features matching all the three models (CV, BR, and DQ), coefficients in the same range, and low computational complexity.

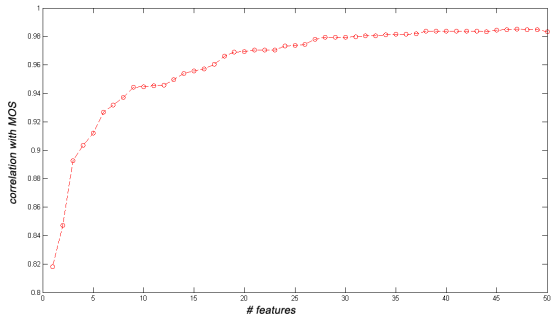


Figure 2: Sequential features selection.

Among the possible combinations, the one presenting at least one coefficient with high magnitude for each model has been chosen following the above mentioned approach, as shown in Table **??**.

Spearman correlation	Features	Execution time $[s]$
0.92	f_{21} f_{22} f_{25} f_{30} f_{32}	211
0.92	f_{11} f_{12} f_{20} f_{25} f_{32}	211
0.92	f_{11} f_{12} f_{25} f_{31} f_{35}	225
0.91	f_{3} f_{4} f_{15} f_{32} f_{36}	256
0.91	f_{11} f_{20} f_{22} f_{25} f_{32}	211
0.91	f_{11} f_{20} f_{22} f_{25} f_{32}	211
0.91	f_{11} f_{12} f_{25} f_{26} f_{27}	408
0.91	f_{15} f_{17} f_{31} f_{36} f_{49}	321
0.91	f_{11} f_{22} f_{30} f_{33} f_{34}	196
0.91	f_{25} f_{32} f_{35} f_{43} f_{48}	238
0.91	f_{3} f_{4} f_{15} f_{32} f_{50}	235
0.91	f_{15} f_{21} f_{22} f_{32} f_{35}	227
0.91	f_{3} f_{4} f_{15} f_{35} f_{36}	276
0.91	f_{15} f_{32} f_{35} f_{43} f_{49}	**238**
...	...	...

Table 4: Sample of 5 features combination.

After the combination of features has been selected, the quality metric is built on those features. In our case the selected ones are **[**f_{15} f_{32} f_{35} f_{43} f_{49}**]**. This metric incorporates all the models: CV (represented by feature f_{15}), BR (represented by features f_{32} and f_{35}), and DQ (represented by features f_{43} and f_{49}). In Figure 3 the overall performances of the proposed algorithm is presented. The fitting of the MOS with the proposed metric shows a good correlation with a value of 0.91.

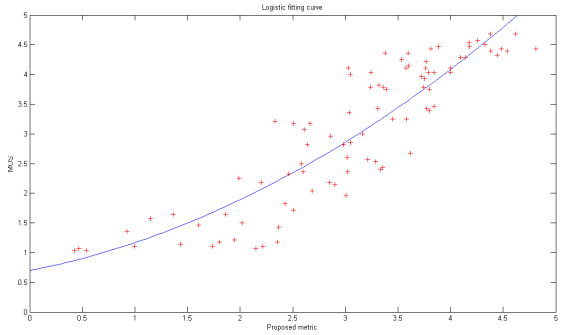


Figure 3: Logistic fitting of the data gathered with the proposed metric.

4. CONCLUSIONS

In this paper a novel stereoscopic metric based on the combination of relevant HVS features has been presented. The quality model has been designed by considering the quality of the cyclopean view, the presence of binocular rivalry, and the presence of binocular depth. Basically, the quality of the single (cyclopean) image obtained by merging the left and right views, the influence of binocular rivalry on visual comfort, and the impact of the depth on correct perception of the 3D scene geometry are considered. To quantify these components, features extracted from the data are analyzed and combined. Here, we select the features whose combination better matches the subjective scores. The collected results show that the proposed metric is able to predict the subjective quality rating in an accurate way while keeping a low computational complexity.

REFERENCES

[1] Schreer, O., Kauff, P., and Sikora, T., [*3D Video Communication Algorithms, concepts and real-time systems in human centered communication*], Wiley (2005).

[2] Allard, J., Franco, J., Menier, C., Boyer, E., and Raffin, B., "The GrImage Platform: A Mixed Reality Environment for Interactions," *Proc. of the IEEE International Conference on Computer Vision Systems (ICVS)* , 46 (2006).

[3] Zilly, F., Muller, M., Eisert, P., and Kauff, P., "The stereoscopic analyzer - an image-based assistance tool for stereo shooting and 3D production," *Proc. of the 17th IEEE International Conference on Image Processing (ICIP)* , 4029 –4032 (2010).

[4] Grau, O., Muller, M., and Kluger, J., "Tools for 3D-TV programme production," *Proc. of the 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)* , 1 –4 (2011).

[5] Boev, A., Hollosi, D., and Gotchev, A., "Classification of stereoscopic artefacts," *MOBILE3DTV* (2012).

[6] Wang, K., Brunnström, K., Barkowsky, M., Urvoy, M., Sjöström, M., Le Callet, P., Tourancheau, S., and Andrén, B., "Stereoscopic 3D video coding quality evaluation with 2D objective metrics," *Proc. SPIE 8648, Stereoscopic Displays and Applications XXIV* **8648**, 86481L–86481L–7 (2013).

[7] Yasakethu, S., Hewage, C., Fernando, W., and Kondoz, A., "Quality analysis for 3D video using 2D video quality models," *IEEE Transactions on Consumer Electronics* **54**(4), 1969–1976 (2008).

[8] Boev, A., Gotchev, A., Egiazarian, K., Aksay, A., and Akar, G., "Towards compound stereo-video quality metric: a specific encoder-based framework," *Proc. of the IEEE Southwest Symposium onImage Analysis and Interpretation* , 218–222 (2006).

[9] Ryu, S., Kim, D. H., and Sohn, K., "Stereoscopic image quality metric based on binocular perception model," *Proc. of the 19th IEEE International Conference on Image Processing (ICIP)* , 609–612 (2012).

[10] Battisti, F., Bosc, E., Carli, M., Le Callet, P., and Perugia, S., "Objective image quality assessment of 3D synthesized views," *Signal Processing: Image Communication* **30**(0), 78 – 88 (2015).

[11] Li, K., Shao, F., Jiang, G., and Yu, M., "Full-reference quality assessment of stereoscopic images by learning sparse monocular and binocular features," *Proc. SPIE 9273, Optoelectronic Imaging and Multimedia Technology III* **9273**, 927312–927312–10 (2014).

[12] Goldmann, L., De Simone, F., and Ebrahimi, T., "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Proc. SPIE 7526, Three-Dimensional Image Processing (3DIP) and Applications* **7526**, 75260S–75260S–11 (2010).

[13] Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricordel, V., Le Callet, P., Gutiérrez, J., and García, N., "Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," *Proc. of the 4th International Workshop on Quality of Multimedia Experience (QoMEX)* , 109–114 (2012).

[14] Jin, L., Boev, A., Egiazarian, K., and Gotchev, A., "Quantifying the importance of cyclopean view and binocular rivalry related features for objective quality assessment of mobile 3D video," *EURASIP Journal on Image and Video Processing, Special Issue on Video Quality Metrics for Consumer Electronics* , 2014:6 (2014).

[15] Baker, S., Roth, S., Scharstein, D., Black, M., Lewis, J., and Szeliski, R., "A database and evaluation methodology for optical flow," *Proc. of IEEE 11th International Conference on Computer Vision (ICCV)* , 1–8 (2007).

[16] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., and Carli, M., "New full-reference quality metrics based on HVS," *Proc. of the Second International Workshop on Video Processing and Quality Metrics* , 9:1 – 9:4 (2006).

[17] Ponomarenko, N., Silvestri, F., Egiazarian, K., Carli, M., and Lukin, V., "On between-coefficient contrast masking of DCT basis functions," *Proc. of the Third International Workshop on Video Processing and Quality Metrics* , 11:1–11:4 (2007).

[18] Zhang, L., Zhang, D., Mou, X., and Zhang, D., "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing* **20**(8), 2378–2386 (2011).

[19] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* **13**(4), 600–612 (2004).