A FULL REFERENCE STEREOSCOPIC VIDEO QUALITY ASSESSMENT METRIC

Balasubramanyam Appina, Manasa K., and Sumohana S. Channappayya

Indian Institute of Technology Hyderabad, Kandi,
India, 502285 e-mail: {ee13m14p100001, ee12p1002, sumohana}@iith.ac.in.

ABSTRACT

We propose a full reference stereo video quality assessment algorithm for assessing the perceptual quality of natural stereo videos. We exploit the separable representation of motion and binocular disparity in the visual cortex and develop a four stage algorithm to measure the quality of a stereoscopic video called FLOSIM$_{3D}$. First, we compute the temporal features by utilizing an existing 2D VQA metric which measures the temporal annoyance based on patch level statistics such as mean, variance and minimum eigen value and pools them with a frame categorization based non-linear pooling strategy. Second, a structure based 2D Image Quality Assessment (IQA) metric is used to compute the spatial quality of the frames. Next, the loss in depth cues is measured using a structure based metric. Finally, the features for each of the stereo views are pooled to obtain the final stereo video quality score. We demonstrate the state-of-the-art performance of the proposed metric on IRCCYN dataset involving H.264, JP2K compression artifacts.

Index Terms— Stereo, full reference video quality assessment, depth, motion.

1. INTRODUCTION

The development and advancements in digital 3D (stereo) multimedia technology has shown rapid growth in the past decade. According to the statistics of Movie Pictures Association of America [1], the 3D movie box office profit reached 11.1 billion USD in 2015 along with a 15% increase in the number of 3D screens compared to 2014. These statistics are a clear indication of the popularity of 3D technology. The source video content can be degraded due to the error prone channel environment leading to a reduction in the Quality of Experience (QoE) of the end user. Objective Quality Assessment (QA) is a useful tool for assessing and improving the user's overall quality of experience. Objective QA is typically classified into full reference (FR), reduced reference (RR) and no reference (NR) based on the utilization of the pristine or reference video content. In this paper, we restrict our focus to FR 3D VQA.

The 3D video quality assessment is more complex compared to 2D VQA due to an additional dimension depth. Early researchers [2, 3, 4], applied the 2D image QA (IQA) metrics by averaging the frame wise scores of left and right views, and 2D video QA (VQA) on the individual views and the final 3D quality score is computed by calculating the mean of both views of the video. Further, they add the depth features to the above quality scores. They concluded that, 2D VQA shows better performance than IQA metrics and the metric performance improved with the inclusion of the depth information. Several studies [5, 6, 7] proposed 3D FR VQA metrics based on local & global patch measurements of luminance, chrominance and depth maps. They computed the statistical and Human Visual System (HVS) inspired measurements on frame by frame basis to compute quality score. They pooled frame quality scores with the weights obtained from depth maps or luminance components to obtain the overall 3D quality score.

Battisti et al. [8] proposed a 3D FR VQA metric based on binocular perceptual properties of HVS. They utilized the cyclopean frames of reference and distorted stereoscopic video, binocular rivalry measurement and binocular depth to perform the QA of stereoscopic video. Jin et al. [9] proposed a stereoscopic FR VQA metric based on 3D-Discrete Cosine Transform (DCT) and Mean Square Error (MSE) computation. They match similar blocks in the right image corresponding to a specific block in the left image and apply 3D-DCT. The final quality score is computed by calculating the MSE between the reference and distorted 3D-DCT coefficients. De Silva et al. [10] proposed a FR 3D VQA based on the measurement of the structural distortion, blur strength and content complexity. The structural distortion strength is computed by calculating the similarity between reference and distorted frames and further, disturbances in edge strength is computed to measure the blur strength. The content complexity is measured by calculating the spatial index (SI) and temporal index (TI) based on ITU recommendation P.910 [11] of a 3D view.

Several studies [12, 13] have proposed 3D VQA metrics by computing the Just Noticeable Differences (JND) from spatial, temporal and depth components of individual views. Han et al. [14] proposed a 3D FR VQA metric based on dependencies in spatial and temporal information of a stereoscopic video. They computed a joint descriptor by calculating the eigenvalues and eigenvectors from 3D tensor structures of a 3D Sobel filtered salient pixel. Kim et al. [15] proposed a

FR 3D VQA metric based on depth map and motion regularities. The spatial quality scores are computed by applying the Scale-Invariant Feature Transform (SIFT) [16] on the frame and temporal consistencies are computed from estimated disparity maps of reference and distorted video sequences.

In this paper, we propose a stereoscopic full reference video quality metric (FLOSIM$_{3D}$) which is based on exploiting the depth and motion dependencies.

2. PROPOSED ALGORITHM

Various studies on the visual cortex revealed that motion and stereoscopic depth are crucial factors to analyse the structure of a visual scene. A visual scene is composed of objects whose position changes over time or across the eyes. The study [17] concluded that a joint-encoding of motion and depth takes place at an early cortical stage. This property of the visual cortex explains the significance of motion and depth to perceive a stereoscopic video. Additionally, it was studied that MT neurons receive a majority of its binocular disparity signals from V2 [18] and the motion information arrives from V1. Therefore, selective pooling of inputs contribute to the largely separable representation of motion and binocular disparity in MT [19]. We exploit this property of the visual cortex and develop a four stage algorithm to measure the quality of a stereoscopic video.

The four-stages of the algorithm are shown in Fig. 1 which involves computing the temporal features, spatial features and depth features. These features for each of the stereo views are pooled to obtain the final stereo video quality score. In the first stage, we extract the temporal quality features from individual views of a stereoscopic video using a motion based 2D-VQA metric. In the second stage, we compute the spatial features using a 2D-IQA metric. The third stage computes the depth maps for each frame of the stereo views. The features from all the three stages are integrated to obtain a frame level score for each of the views. The frame-level scores are pooled to view-wise scores and the two view-wise scores are averaged to obtain a stereo video quality score.

2.1. Temporal Features

Binocular retina of the HVS captures two different views of a single scene point to create an illusion of depth perception. DeAngelis and Newsome [20] showed that, the disparity selective neurons present in the area MT have functional organization of patch based dependencies between motion and disparity. Additionally, the perceptual depth quality can be affected by the motion information due to the dependencies on dynamic and kinetic depth cues, motion parallax [21, 22] of a 3D view and also, several studies [2, 23] explored the relationship between motion variation and depth quality based on perceptual distortion strengths. Dufaux et al. [24] explored the motion and depth dependencies on structural properties

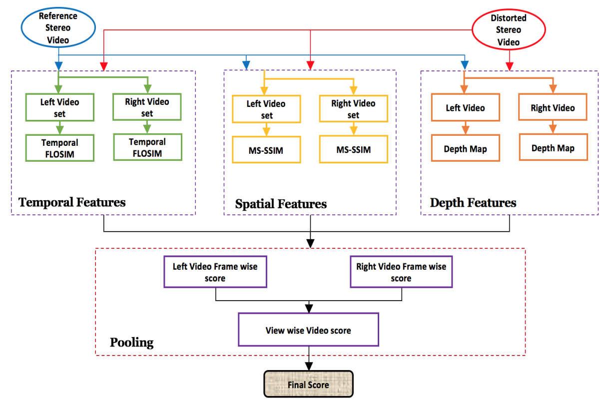


Fig. 1: Overview of FLOSIM$_{3D}$.

of an edge in a 3D scene. They concluded that the compression and distorted artifacts change the structural properties of an edge and it varies the motion and depth perception compared to the reference video. This implies that frame categorization based on disturbances in temporal information due to artifacts can give a better utilization of depth quality features. Therefore, the selection of the 2D VQA metric to compute the temporal features should imbibe these properties. We chose to work with FLOSIM [25] which is based on measuring the patch level statistical properties of the optical flow (finest representation of motion information). Additionally, FLOSIM pools the temporal features using a non-linear pooling strategy by categorizing the frames which gives better idea on the effect of distortion on disparity and motion variation between frames of a 3D video.

The temporal FLOSIM measures the local flow statistics with patch wise mean, μ, variance, σ and minimum eigenvalue, λ_{min}. The difference in dispersion of the mean and variance of all the patches in a distorted frame and the reference frame is measured. The patch-wise correspondence of each patch is measured by taking the correlation of the minimum eigen value of the reference and distorted patches. Equations (1) and (2) give the frame level temporal distortion measure.

$$D^i_{\mathbf{x}}(\mathbf{V}^i_p, \mathbf{V}^i_t) = CV_{\mathbf{x}}(\mathbf{V}^i_p) - CV_{\mathbf{x}}(\mathbf{V}^i_t), \quad (1)$$

where **x** = $\{$left(l), right(r)$\}$, **V**= $\{\mu_{|.|}\ , \sigma_{|.|}\}$, $D^i_{\mathbf{x}}(\mathbf{V}^i_p, \mathbf{V}^i_{t_f})$ is the dispersion difference of the frame i in the **x** view (left or right), of the p pristine and t test videos of the feature **V** (mean or variance), $CV(\mathbf{V}^i_t)$ is the coefficient of variation of the feature **V** of the frame i for the test video t, $\mathbf{V}_i$ represents the features $(\mu_{|.|}, \sigma_{|.|})$ of all the patches in the frame i.

The minimum eigen value of the covariance matrix of horizontal and vertical flow components of a flow patch is a measure of randomness in the flow patch. The frame level randomness is measured by

$$C^i_{\mathbf{x}}(\mathbf{V}^i_p, \mathbf{V}^i_t) = 1 - \text{corr}_{\mathbf{x}}(\mathbf{V}^i_p, \mathbf{V}^i_t), \quad (2)$$

where $\mathbf{V}$= $\{\lambda_{|.|}\}$, and $\text{corr}_{\mathbf{x}}$ is the linear correlation coefficient.

The frame level distortion is however not a representative of the video quality as the distortions might go unseen in a video. To measure the perceptual annoyance of the video, the frame categorisation is done with two thresholds τ_{μ}^{i}, τ_{σ}^{i} which are based on temporal features of the neighbouring frames. Table 1 tabulates the four categories of regions and their corresponding temporal FLOSIM score (Q_{FL}^{i}) and also explains the motion and depth based prominence of each region.

The temporal annoyance levels are measured on individual videos of a stereoscopic view and $Q_{FL_l}^{i}$, $Q_{FL_r}^{i}$ represents the left and right video temporal FLOSIM scores. The overall temporal FLOSIM score of stereoscopic video is computed as

$$Q_{FL} = \frac{1}{2T}\sum_{i=1}^{T}(Q_{FL_l}^{i} + Q_{FL_r}^{i}), \tag{3}$$

where T represents the total number of frames.

2.2. Spatial Features

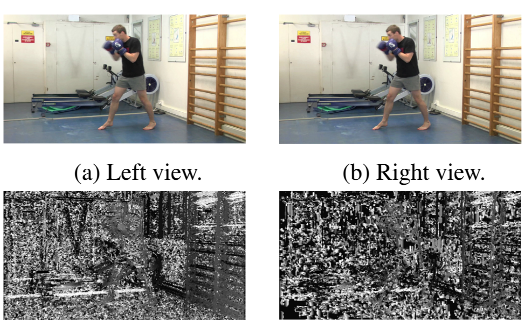
The HVS is highly sensitive and is adapted to the structural properties [28] of a view. MS-SSIM [29] is a top down approach which computes the structural similarity at different scales and hence, the spatial quality for each frame of the left and the right view is given by

$$Q_{s_l}^{i} = (1 - (\text{MS-SSIM}(H_{s_{p_l}}^{i}, H_{s_{t_l}}^{i})))$$
$$Q_{s_r}^{i} = (1 - (\text{MS-SSIM}(H_{s_{p_r}}^{i}, H_{s_{t_r}}^{i})))$$

where, $H_{s_{p_l}}^{i}$ represents the frame i of the pristine p video of the left view l. s is a symbol for spatial content.

2.3. Depth Features

The retinal views of both the eyes converge to a point in the visual space to illustrate the single projection of the perceived scene. The simple and complex cells extracts the structural information (spatial and temporal) of both the eyes and this information is processed in primary visual cortex to make an illustration of depth perception [30].



(a) Left view. (b) Right view.

(c) Reference disparity. (d) Distorted disparity.

Fig. 2: Luminance and disparity maps.

In order to achieve this, we chose SSIM based stereo matching algorithm [31] based on trade-off between accuracy and time complexity. Fig. 2 shows a frame of left & right views and the corresponding depth map of reference and distorted views (QP = 44).

In our approach, we computed two disparity maps from a stereo view by treating the left view as reference (d_l^i) and next by treating right view as reference (d_r^i). Therefore, two disparity maps are computed for a reference stereo video ($d_{p_l}^i$, $d_{p_r}^i$) and a distorted stereo video ($d_{t_l}^i$, $d_{t_r}^i$). In our work, we resort to quantifying the loss in depth cues of distorted video compared to the reference video. The depth perception in the HVS is highly sensitive to the structural properties [30, 32] of a scene point. Therefore, the loss in depth cue is measured by computing the structural information loss in distorted video depth maps with respect to the reference video depth maps [33] with MS-SSIM.

$$Q_{d_l}^{i} = (1 - \text{MS-SSIM}(d_{p_l}^{i}, d_{t_l}^{i})), \tag{4}$$

$$Q_{d_r}^{i} = (1 - \text{MS-SSIM}(d_{p_r}^{i}, d_{t_r}^{i})), \tag{5}$$

where $Q_{d_l}^{i}$ and $Q_{d_r}^{i}$ measure the depth quality of left and right views respectively. The overall depth quality score is computed by taking the mean of Q_{d_l}, Q_{d_r} across all the frames.

$$Q_d = \frac{1}{2T}\sum_{i=1}^{T}(Q_{d_l}^{i} + Q_{d_r}^{i}). \tag{6}$$

2.4. Overall Quality Score

To compute the overall quality of a stereoscopic video, first we pooled the temporal and spatial features into a single score by computing the mean of product of individual features of each view across frames according to (7) and further, we integrated the depth quality features by (8).

$$\text{FLOSIM} = \frac{1}{2T}\sum_{i=1}^{T}(Q_{s_l}^{i}Q_{FL_l}^{i} + Q_{s_r}^{i}Q_{FL_r}^{i}), \tag{7}$$

$$\text{FLOSIM}_{3D} = \text{FLOSIM} \times Q_d. \tag{8}$$

We are motivated from [36] to give equal priority to all components.

3. RESULTS AND DISCUSSION

The performance of the proposed metric FLOSIM_{3D} was tested on IRCCYN dataset [37]. Due to the lack of publicly available distorted 3D video datasets, we limited our evaluation to IRCCYN dataset. The IRCCYN dataset has 10 reference video sequences with good variety of structure, texture, depth and temporal information. These video are captured with Panasonic AG-3DA1E twin-lens camera with

Table 1: Frame categorisation based on perceptual annoyance levels and prominence in motion and depth regime

Region	Threshold	Significance in motion	Significance in depth	Temporal FLOSIM (Q_{FL}^i)
R1	$D_{\mathbf{x}}^i(\boldsymbol{\mu}_p^i, \boldsymbol{\mu}_t^i) > \tau_\mu^i$, $D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i) > \tau_\sigma^i$	Higher irregularities in flow field	Non-uniform and inconsistencies in depth map [26]	$D_{\mathbf{x}}^i(\boldsymbol{\mu}_p^i, \boldsymbol{\mu}_t^i) + D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i).C_{\mathbf{x}}^i(\boldsymbol{\lambda}_p^i, \boldsymbol{\lambda}_t^i)$
R2	$D_{\mathbf{x}}^i(\boldsymbol{\mu}_p^i, \boldsymbol{\mu}_t^i) < \tau_\mu^i$, $D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i) > \tau_\sigma^i$	Higher intra patch irregularities	Motion parallax and rivalry errors in depth [27]	$D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i).C_{\mathbf{x}}^i(\boldsymbol{\lambda}_p^i, \lambda_t^i)$
R3	$D_{\mathbf{x}}^i(\boldsymbol{\mu}_p^i, \boldsymbol{\mu}_t^i) < \tau_\mu^i$, $D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i) < \tau_\sigma^i$	Acceptable temporal distortion	Tolerable depth errors	$C_{\mathbf{x}}^i(\boldsymbol{\lambda}_p^i, \boldsymbol{\lambda}_t^i)$
R4	$D_{\mathbf{x}}^i(\boldsymbol{\mu}_p^i, \boldsymbol{\mu}_t^i) > \tau_\mu^i$, $D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i) < \tau_\sigma^i$	Higher inter patch irregularities	Dynamic occlusion errors in depth [27]	$D_{\mathbf{x}}^i(\boldsymbol{\sigma}_p^i, \boldsymbol{\sigma}_t^i).C_{\mathbf{x}}^i(\boldsymbol{\lambda}_p^i, \boldsymbol{\lambda}_t^i)$

Table 2: 2D & 3D I/VQA performance evaluation on IRCCYN dataset.

Algorithm	**H.264**			**JP2K**			**Overall**		
	LCC	**SROCC**	**RMSE**	**LCC**	**SROCC**	**RMSE**	**LCC**	**SROCC**	**RMSE**
SSIM	0.6223	0.5464	0.8843	0.7137	0.5974	0.9202	0.5598	0.2465	1.0264
MS-SSIM	0.7885	0.6673	0.6955	0.9439	0.9299	0.4327	0.8506	0.8534	0.6512
Chen [31]	0.662	0.5720	0.6915	0.8817	0.8724	0.6182	0.798	0.7861	0.7464
STRIQE [34]	0.7913	0.7167	0.8433	0.9017	0.8175	0.5666	0.7931	0.7734	0.7544
STMAD [35]	0.7641	0.7354	0.7296	0.8388	0.7236	0.7136	0.6400	0.3495	0.9518
Q_{FL}	0.6453	0.5489	0.6958	0.8441	0.8278	0.7027	0.7252	0.7097	0.8528
FLOSIM [25]	0.9265	0.8987	0.4256	0.9665	0.9495	0.3359	0.9074	0.8986	0.5206
Q_d	0.5641	0.5181	0.9571	0.6466	0.5537	0.9997	0.6167	0.5759	0.9751
Chen$_{3D}$	0.7963	0.8035	2.5835	0.9358	0.8884	3.2863	0.8227	0.8201	2.9763
STRIQE$_{3D}$	0.6836	0.6263	2.3683	0.8778	0.8513	3.2121	0.7599	0.7525	2.8374
FLOSIM$_{\mathbf{3D}}$(proposed)	**0.9589**	**0.9478**	**0.3863**	**0.9738**	**0.9548**	**0.2976**	**0.9178**	**0.9111**	**0.4918**

a baseline seperation of 60 mm. The video sequences have a resolution of 1920 $\times$ 1080 and duration of either 16 seconds or 13 seconds with a frame speed of 25 fps. The H.264 compression artifacts are introduced using JM reference software by varying the quantization parameter (QP = 32,38,44) and JP2K artifacts (2,8,16,32 Mb/s) are applied on frame basis of a view. These artifacts are applied symmetrically on left and right videos.

We report the Linear Correlation Coefficient (LCC), Spearmans Rank order correlation (SROCC) and Root Mean Square Error (RMSE) scores between different 2D & 3D I/VQA metrics and DMOS scores after non-linear logistic fit [38]. Here, higher SROCC and LCC scores indicate good agreement with HVS scores and lower RMSE indicates better performance. Also, we extended several 3D IQA metrics [31, 34] to 3D VQA (Chen$_{3D}$, STRIQE$_{3D}$) metrics by including the temporal features (Q_{FL}) for better comparison. The computation of Chen$_{3D}$, STRIQE$_{3D}$ is shown below:

$$\text{Chen}_{3D} = \frac{1}{T}\sum_{i=1}^{T}\left(MJ_{3D}^i \times \left[\frac{Q_{FL_l}^i + Q_{FL_r}^i}{2}\right]\right),$$

$$\text{STRIQE}_{3D} = \frac{1}{T}\sum_{i=1}^{T}\left(ST_{3D}^i \times \left[\frac{[Q_{FL_l}^i + Q_{FL_r}^i}{2}\right]\right),$$

where MJ_{3D}^i is the spatial quality score of frame pair i as computed using [31], and ST_{3D}^i is the spatial quality score of frame pair i as computed using [34].

Table 2 shows the performance evaluation of different 2D IQA/VQA and 3D IQA metrics on the IRCCYN dataset. It is clear that FLOSIM$_{3D}$ outperforms the other 2D IQA/VQA and 3D IQA metrics. Even though, the video sequences have been distorted at different compression rates, they do not have significant perceptual distortion. Due to this, 2D IQA metrics show competitive performance with 3D IQA metrics.

4. CONCLUSION

A FR 3D VQA metric called FLOSIM$_{3D}$ was proposed based on the dependencies between motion and its disparities. We utilized the frame categorization of a video sequence by using optical flow field strength to compute the dependencies. Further, we included the depth features in addition to spatial and temporal features to compute quality. We showed that FLOSIM$_{3D}$ has state-of-the-art performance on the IRCCYN stereoscopic video dataset and easily outperforms existing 2D/3D I/VQA metrics. In future, we plan to improve and extend the metric performance by including features of aforementioned dependencies and evaluate its performance on a wider variety of 3D videos.

5. REFERENCES

[1] Motion Picture Association of America, "Theatrical market statistics 2013," 2015.

[2] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, "Quality analysis for 3d video using 2d video quality models,"

IEEE Transactions on Consumer Electronics, vol. 54, pp. 1969–1976, November 2008.

[3] 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*, vol. 44, pp. 963–965, July 2008.

[4] C. D. M. Regis, J. V. de Miranda Cardoso, . de Pontes Oliveira, and M. S. de Alencar, "Objective estimation of 3d video quality: A disparity-based weighting strategy," in *2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, June 2013.

[5] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando, and A. M. Kondoz, "Perceptual video quality metric for 3d video quality assessment," in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 1–4, June 2010.

[6] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos, "3d video quality metric for 3d video compression," in *IVMSP Workshop, 2013 IEEE 11th*, pp. 1–4, June 2013.

[7] C. Sun, X. Liu, X. Xu, and W. Yang, "An efficient quality assessment metric for 3d video," in *Computer and Information Technology (CIT), 2012 IEEE 12th International Conference on*, pp. 209–213, Oct 2012.

[8] F. Battisti, M. Carli, A. Stramacci, A. Boev, and A. Gotchev, "A perceptual quality metric for high-definition stereoscopic 3d video," in *SPIE/IS&T Electronic Imaging*, pp. 939916–939916, International Society for Optics and Photonics, 2015.

[9] L. Jin, A. Gotchev, A. Boev, and K. Egiazarian, "Validation of a new full reference metric for quality assessment of mobile 3dtv content," in *Signal Processing Conference, 2011 19th European*, pp. 1894–1898, Aug 2011.

[10] V. D. Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video," *IEEE Transactions on Image Processing*, vol. 22, pp. 3392–3404, Sept 2013.

[11] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[12] F. Qi, T. Jiang, X. Fan, S. Ma, and D. Zhao, "Stereoscopic video quality assessment based on stereo just-noticeable difference model," in *2013 IEEE International Conference on Image Processing*, pp. 34–38, Sept 2013.

[13] D. V. S. X. D. Silva, W. A. C. Fernando, G. Nur, E. Ekmekcioglu, and S. T. Worrall, "3d video assessment with just noticeable difference in depth evaluation," in *2010 IEEE International Conference on Image Processing*, pp. 4013–4016, Sept 2010.

[14] J. Han, T. Jiang, and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," in *Visual Communications and Image Processing (VCIP), 2012 IEEE*, pp. 1–6, Nov 2012.

[15] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn, "Depth map quality metric for three-dimensional video," in *IS&T/SPIE Electronic Imaging*, pp. 723719–723719, International Society for Optics and Photonics, 2009.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

[17] A. Anzai, I. Ohzawa, and R. D. Freeman, "Joint-encoding of motion and depth by visual cortical neurons: neural basis of the pulfrich effect," *Nature neuroscience*, vol. 4, no. 5, pp. 513–518, 2001.

[18] B. Cumming and G. DeAngelis, "The physiology of stereopsis," *Annual review of neuroscience*, vol. 24, no. 1, pp. 203–238, 2001.

[19] A. Smolyanskaya, D. A. Ruff, and R. T. Born, "Joint tuning for direction of motion and binocular disparity in macaque mt is largely separable," *Journal of neurophysiology*, vol. 110, no. 12, pp. 2806–2816, 2013.

[20] G. C. DeAngelis and W. T. Newsome, "Organization of disparity-selective neurons in macaque area mt," *The Journal of neuroscience*, vol. 19, no. 4, pp. 1398–1415, 1999.

[21] B. T. Backus, M. S. Banks, R. van Ee, and J. A. Crowell, "Horizontal and vertical disparity, eye position, and stereoscopic slant perception," *Vision research*, vol. 39, no. 6, pp. 1143–1170, 1999.

[22] R. S. Harwerth, P. M. Fredenburg, and E. L. Smith, "Temporal integration for stereoscopic vision," *Vision research*, vol. 43, no. 5, pp. 505–517, 2003.

[23] K. Ha and M. Kim, "A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video," in *2011 18th IEEE International Conference on Image Processing*, pp. 2525–2528, IEEE, 2011.

[24] F. Dufaux, B. Pesquet-Popescu, and M. Cagnazzo, *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*. Wiley, 2013.

[25] M. K. and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, pp. 2480–2492, June 2016.

[26] A. D. Hwang and E. Peli, "Instability of the perceived world while watching 3d stereoscopic imagery: A likely source of motion sickness symptoms," *i-Perception*, vol. 5, no. 6, pp. 515–535, 2014.

[27] R. Patterson, *Human Factors of Stereoscopic 3D Displays*. Springer London, 2015.

[28] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *Journal of neurophysiology*, vol. 28, no. 2, pp. 229–289, 1965.

[29] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, pp. 1398–1402, Ieee, 2003.

[30] W. J. Levelt, *On binocular rivalry*, vol. 2. Mouton The Hague, 1968.

[31] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.

[32] B. Appina, S. Khan, and S. S. Channappayya, "No-reference stereoscopic image quality assessment using natural scene statistics," *Signal Processing: Image Communication*, vol. 43, pp. 1–14, 2016.

[33] R. R. Tamboli, B. Appina, S. Channappayya, and S. Jana, "Super-multiview content with high angular resolution: 3d quality assessment on horizontal-parallax lightfield display," *Signal Processing: Image Communication*, vol. 47, pp. 42–55, 2016.

[34] S. Khan Md, B. Appina, and S. Channappayya, "Full-reference stereo image quality assessment using natural stereo scene statistics," *Signal Processing Letters, IEEE*, vol. 22, pp. 1985–1989, Nov 2015.

[35] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *2011 18th IEEE International Conference on Image Processing*, pp. 2505–2508, IEEE, 2011.

[36] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010.

[37] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, and N. Garcia, "Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pp. 109–114, IEEE, 2012.

[38] "Vqeg. (aug. 2003). final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii. [online]. available:
http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx."