NO-REFERENCE STEREOSCOPIC VIDEO QUALITY ASSESSMENT ALGORITHM USING JOINT MOTION AND DEPTH STATISTICS

Balasubramanyam Appina, Akshith Jalli, Shanmukha Srinivas Battula, and Sumohana S. Channappayya

Indian Institute of Technology Hyderabad, Kandi, India, 502285
e-mail: {ee13m14p100001, ee13b1013, ee13b1006, sumohana}@iith.ac.in.

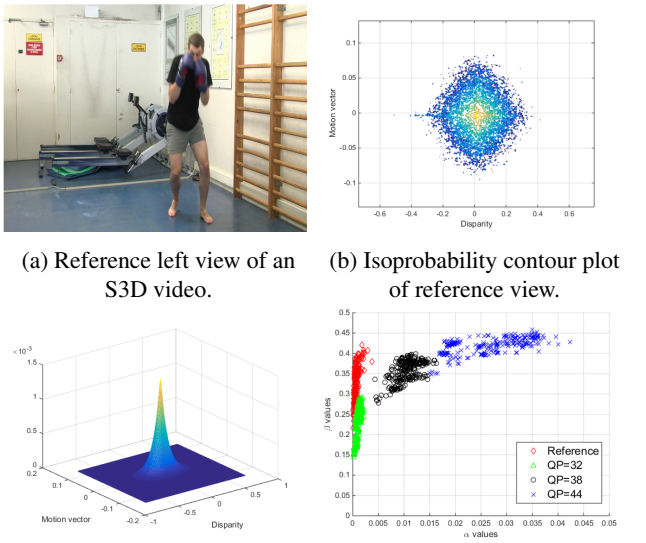
ABSTRACT

We propose a supervised no-reference (NR) quality assessment algorithm for assessing the perceptual quality of natural stereoscopic (S3D) videos. We empirically model the joint statistics of motion and depth subband coefficients of an S3D video frame using a Bivaraite Generalized Gaussian Distribution (BGGD). We compute the BGGD model parameters (α, β) to estimate the statistical dependency strength and show the features are quality discriminative. We compute the popular 2D NR image quality assessment (IQA) model NIQE on a frame-by-frame basis for both views to estimate the spatial quality. The frame-level BGGD features and spatial features are consolidated and used with the corresponding S3D videos difference mean opinion score (DMOS) labels for supervised learning using support vector regression (SVR). The overall quality of an S3D video is computed by averaging the frame-level quality predictions of the constituent video frames. The proposed algorithm, dubbed Video QUality Evaluation using MOtion and DEpth Statistics (VQUEMODES) is shown to outperform the state-of-the-art methods when evaluated over the IRCCYN and LFOVIA S3D subjective quality assessment databases.

Index Terms— Stereoscopic Video, Perceptual Quality, Natural Scene Statistics.

1. INTRODUCTION

In this work, we propose a supervised NR video quality assessment (VQA) model for S3D videos. Several S3D NR VQA models [1, 2, 3, 4] have been proposed based on spatiotemporal segmentation, spatial structural loss measurement, motion inconsistencies, etc. More recently, Yang *et al.* [5] proposed an S3D supervised NR VQA metric based on a multi view binocular perception model. They applied the curvelet transform on spatial information of an S3D video to extract the texture analysis features and optical flow features were utilized to measure the temporal quality. Finally, they used empirical weight combinations to pool these scores to compute the overall quality score. Ha and Kim [6] proposed an S3D NR VQA metric based on temporal variance, intra and inter disparity measurements. Depth maps are computed by minimizing the mean square error values, and motion vector length is calculated to estimate the temporal variations. Intra and inter frame disparities were computed to measure the dependencies between motion and depth components. Chen *et al.* [7] proposed an S3D NR VQA model based on a binocular energy mechanism. They computed the auto-regressive prediction based disparity measurement and rely on natural scene statistics of an S3D video to compute the quality. Jiang *et al.* [8] proposed an S3D NR supervised VQA model based on tensor decomposed motion feature extraction. They computed the univariate GGD, asymmetric GGD, spatial and spectral entropy from the tensor decomposition. The random forest classifier was used to predict the quality of an S3D video.



(a) Reference left view of an S3D video.

(b) Isoprobability contour plot of reference view.

(c) Bivariate GGD fit of reference view. $\chi = 2 \times 10^{-7}$.

(d) α an d β parameter variation at different H.264 distortion levels.

Fig. 1: Illustration of isoprobability contours and BGGD fits between motion vector and disparity map statistics of pristine Boxers S3D video of IRCCYN database. BGGD model parameter variation of reference and H.264 distortion levels of corresponding S3D video.

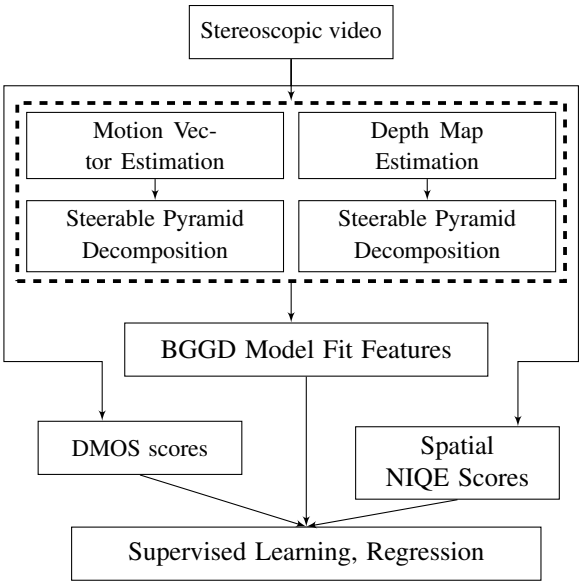


Fig. 2: Flowchart of the proposed VQUEMODES algorithm.

None of the above algorithms have studied or utilized the statistical dependencies between motion and depth components. We propose an NR VQA algorithm for S3D videos based on quantifying the statistical dependency between the motion and depth components of an S3D video combined with a spatial quality estimate. Our algorithm is called Video QUality Evaluation using MOtion and DEpth Statistics (VQUEMODES) and is described next.

2. PROPOSED ALGORITHM

Psychovisual experiments on the mammalian visual cortex to explore the disparity selectivity in middle temporal (MT) area of the brain [9, 10] have concluded that the MT neurons are not responsible for motion processing but are also highly tuned for binocular disparity processing. Motivated by these findings, we attempt to model the joint statistical dependencies between motion and depth statistics using a Bivariate Generalized Gaussian Distribution (BGGD).

While this model is inspired from our work in [11], we would like to highlight that a statistical analysis of the dependencies between motion and depth subband coefficients of natural S3D videos has not been carried out previously (to the best of our knowledge). Through this work we also highlight the utility of this model in an NR VQA application. We model the joint statistics of motion and depth components using a BGGD and estimate the BGGD model parameters (α, β) to quantify the statistical dependency between the motion and depth components of an S3D video frame. We show the features are distortion discriminable and use them in the quality computation of an S3D video.

2.1. BGGD Modeling

We empirically show that a BGGD accurately models the joint histogram of motion and depth subband coefficients of an S3D view. Consider a multivariate GGD distribution of a random vector $\mathbf{x} \in \mathbb{R}^N$ given as [11]

$$p(\mathbf{x}|\mathbf{M},\alpha,\beta) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\alpha,\beta}(\mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}),$$

$$g_{\alpha,\beta}(y) = \frac{\beta\Gamma(\frac{N}{2})}{(2^{\frac{1}{\beta}}\Pi\alpha)^{\frac{N}{2}}\Gamma(\frac{N}{2\beta})} e^{-\frac{1}{2}(\frac{y}{\alpha})^{\beta}},$$

where $\mathbf{M}$ is an $N \times N$ symmetric covariance matrix and $g_{\alpha,\beta}(\cdot)$ is the density generator. Since motion and depth are the two parameters in the model, $N = 2$. Therefore, the above multivariate GGD becomes a bivariate GGD. We computed the α and β scores at multiple scales (3 scales) and multiple orientations ($0^0, 30^0, 60^0, 90^0, 120^0, 150^0$) using the steerable pyramid decomposition [12].

Fig. 1a shows the 100^{th} frame from the left view of the pristine Boxers S3D video from the IRCCYN database [13]. Fig. 1b shows the isoprobability contour plot of joint subband coefficients of the motion vector and depth map of an S3D video frame and Fig. 1c shows the estimated BGGD fit of respective contour plot Fig. 1b. The motion vectors are computed using the standard three-step search method [14]. The disparity map is computed using the SSIM based stereo matching algorithm [15]. The contour plot clearly shows the dependencies between motion and depth components. Specifically, we arrive at this conclusion due to the circular asymmetry in the plots. We have further verified that the product of the marginals is different from the joint distributions.

The dependencies between motion and depth components vary with distortion and is reflected by the changes in the model parameters (α, β). Fig. 1d shows the BGGD features of the pristine Boxers S3D video and its H.264 compressed versions. It is clear that the estimated BGGD features are well segregated with respect to the perceptual quality level. Also, we check the efficacy of proposed model by computing the goodness of fit (χ) value between estimated fit and our observation. In our analysis we observed that χ are in the range 10^{-8} and 10^{-6} for all S3D video sequences. These observations motivate us to use the BGGD features in the quality computation of an S3D video. The plots in Fig. 1 are shown for the first scale and 0^0 orientation of the steerable pyramid decomposition.

2.2. No-Reference Video Quality Assessment

The flowchart of the proposed algorithm is shown in Fig. 2. The feature extraction stage estimates frame-wise BGGD model parameters to represent motion and depth quality features, and relies on the NIQE score [16] as the spatial quality feature. These features are then used to train an SVR for frame-wise quality score estimation. For video-level quality

Table 1: 2D & 3D I/VQA performance evaluation on the IRCCYN S3D video database.

Model Type	Algorithm	H.264			JP2K			Overall		
		LCC	**SROCC**	**RMSE**	**LCC**	**SROCC**	**RMSE**	**LCC**	**SROCC**	**RMSE**
2D NR IQA	BRISQUE [17]	0.7915	0.7637	0.7912	0.8048	0.8999	0.5687	0.7535	0.8145	0.6535
	NIQE [16]	0.6403	0.6617	0.8686	0.8808	0.7240	0.6206	0.5524	0.4183	1.0326
3D FR IQA	Chen *et al.* [15]	0.6620	0.5720	0.6915	0.8817	0.8724	0.6182	0.7980	0.7861	0.7464
	STRIQE [18]	0.7913	0.7167	0.8433	0.9017	0.8175	0.5666	0.7931	0.7734	0.7544
3D FR VQA	FLOSIM_{3D} [19]	0.9589	0.9478	0.3863	0.9738	0.9548	0.2976	0.9178	0.9111	0.4918
	PQM [20]	-	-	-	-	-	-	0.6340	0.6006	0.8784
	PHVS-3D [21]	-	-	-	-	-	-	0.5480	0.5146	0.9501
	3D-STS [22]	-	-	-	-	-	-	0.6417	0.6214	0.9067
	SJND-SVA [23]	0.5834	0.6810	0.6672	0.8062	0.6901	0.5079	0.6503	0.6229	0.8629
	3-D-PQI [24]	0.9306	0.9239	-	0.9413	0.9266	-	0.9009	0.8848	-
	DeMo_{3D} [25]	0.9161	0.9009	0.4564	0.9505	0.9326	0.4074	0.9272	0.9187	0.4651
3D NR VQA Supervised	Yang *et al.* [5]	-	-	-	-	-	-	0.8949	0.8552	0.4929
	BSVQE [7]	0.9168	0.8857	-	0.8953	0.8383	-	0.9239	0.9086	-
	MNSVQM [8]	0.8850	0.7714	0.4675	0.9706	0.8982	0.2769	0.8611	0.8394	0.5634
	BGGD features only	0.9253	0.8955	0.3555	0.9690	0.9477	0.2572	0.9569	0.9330	0.3162
	VQUEMODES (NIQE)	**0.9594**	**0.9439**	**0.1791**	**0.9859**	**0.9666**	**0.0912**	**0.9697**	**0.9637**	**0.2635**

prediction, the individual frame-wise quality predictions are simply averaged. The algorithm is described in detail in the following.

2.2.1. Feature Extraction

- Motion and Depth Features: Three-step motion vector estimation method [14] is used to compute the motion vector map at a macroblock size of 8×8. The magnitude of motion vector is computed and utilized as the motion feature in our algorithm. An SSIM based stereo matching algorithm [15] is used to compute the disparity maps in our algorithm.

  Steerable pyramid decomposition was performed on the estimated motion vector and disparity maps at multiple scales and orientations. To maintain the consistency with the motion vector block size of 8×8 the depth maps are averaged to match this block size.

- BGGD model parameter estimation: As mentioned previously, we used three spatial scales and six orientations in our analysis resulting in a total of 18 subbands for every stereoscopic video frame. The BGGD model parameters are computed at every subband resulting in a feature vector $f = [\alpha^1 \dots \alpha^{18}; \beta^1 \dots \beta^{18}]$ per frame. For an S3D video, the feature vector set is $[f_1, f_2 \dots f_n]$, where n is the number of video frames and $f_i = [\alpha_i^1 \dots \alpha_i^{18}; \beta_i^1 \dots \beta_i^{18}]; 1 \leq i \leq n-1$.

- Spatial Feature: We evaluate the NIQE [16] model on the frame-by-frame basis of each view of an S3D video to compute the spatial feature. NIQE is an opinion and distortion unaware NR 2D IQA model.

$$S = \frac{1}{n} \times \sum_{i=1}^{n} \frac{NIQE_i^L + NIQE_i^R}{2},$$

where L, R represent the left and right views of an S3D video. n indicates the total number of frames of an S3D video. $NIQE^L$ and $NIQE^R$ represent the frame level NIQE scores of the left and right views. S represents the overall spatial quality of an S3D video.

2.3. Supervised Learning and Quality Estimation

The spatial quality feature (S_i) is appended to the aforementioned BGGD features to form the feature vector of a video frame $f_i^s = [\alpha_i^1 \dots \alpha_i^{18}; \beta_i^1 \dots \beta_i^{18}; S_i]$. We believe that over short temporal durations, the average DMOS score of an S3D video and the frame-level DMOS score are highly correlated and are interchangeable. Therefore, we performed the regression of the frame-level features f_i^s and the video-level DMOS scores D as its label. For video V,

$$f_i^{s^V} = [\alpha_i^1 \dots \alpha_i^{18}; \beta_i^1 \dots \beta_i^{18}; S_i],$$

with the corresponding label D_V. This feature vector and label set is used to train an SVR. SVR is shown to provide good performance even when the available training set size is small, demonstrate accurate performance in one-versus-rest schemes, provide sparse solutions and accurate estimation of global minimum etc. In our work, we used the radial basis function (RBF) kernel as it gave the best overall performance.

We use regression to estimate the scores of test video frames. It should be noted that the training and regression

happen at the frame-level. The overall (video-level) quality score is estimated by averaging the frame-level quality estimates.

3. RESULTS AND DISCUSSION

The efficacy of proposed algorithm is evaluated on the IRCCYN [13] and the LFOVIA [26] S3D video databases.

The IRCCYN database has 10 reference and 70 test S3D video sequences. The video sequences have a resolution of 1920×1080 and saved in .avi container. The frame rate is 25 fps and a duration of either 16 sec or 13 sec. The database is a combination of H.264 (QP=32, 38, 44) and JP2K (Bitrate = 2, 8, 16, 32 Mb/s) distorted S3D video sequences. These artifacts are applied symmetrically on each view of an S3D video and published the DMOS scores as subjective scores. The LFOVIA database [26] has H.264 compressed stereoscopic video sequences. The database has 6 pristine and 144 distorted video sequences and saved in .mp4 container. The video sequences have a resolution of 1836×1056 pixels with a frame rate of 25 fps and a duration of 10 sec. They used four different bitrates (100, 200, 350, 1200 Kbps) and created 24 symmetric and 120 asymmetric distorted S3D videos. The subjective study is performed using the ACR-HR method and published DMOS scores as subjective scores.

For both the databases, 80% of the videos are used for SVR training and the remaining samples are used for regression. In other words, the training and test sets are obtained by partitioning the set of available videos in the 80:20 proportion. Once this video-level partitioning is done, the actual training happens at the frame-level. During regression, the frame-level scores are estimated and averaged to compute the video-level quality score. We empirically justify the averaging of the frame-level scores to generate the video-level score. In over 1000 regression iterations, we found that the standard deviation of frame-level scores for a given video varied between 0.2×10^{-8} and 0.25.

We used the open-source SVM package *LIBSVM* [27] in our experiments. We performed the training and testing 1000 times for statistical consistency with a random assignment of video-level samples without overlap between the training and testing sets. The reported results are the average over these 1000 trials. The performance of the proposed metric is measured using the following statistical measures: Linear Correlation Coefficient (LCC), Spearman's Rank Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE). All these results are evaluated after performing a non-liner logistic fit [28].

Tables 1 and 2 shows the performance evaluation of proposed metric on IRCCYN [13] and LFOVIA [26] S3D video databases. Also, we compared the proposed metric results with different popular 2D and 3D IQA/VQA metric performances. BRISQUE [17] and NIQE [16] are 2D NR IQA models. Chen *et al.* [15] and STRIQE [18] are 3D FR IQA metrics. These IQA metrics were applied on a frame-by-frame basis for each view, and the final S3D quality score is computed by calculating the mean score of all frame scores of both views. FLOSIM_{3D} [19], PQM [20], PHVS-3D [21], 3D-STS [22], SJND-SVA [23], 3-D-PQI [24] and DeMo_{3D} [25] are popular S3D FR VQA models. Yang *et al.* [5], BSVQE [7] and MNSVQM [8] are S3D NR VQA models. From the results, it is clear that the proposed metric out performs all of the 2D and 3D IQA/VQA FR and NR models on IRCCYN and LFOVIA S3D video databases.

Table 2: Performance evaluation on LFOVIA S3D Video Database.

Algorithm	LCC	SROCC	RMSE
NIQE [16]	0.7206	0.7376	11.1138
STMAD [5]	0.6802	0.6014	9.4918
DeMo_{3D} [25]	0.9033	0.8991	5.0392
VQUEMODES (NIQE)	**0.8943**	**0.8890**	**5.9124**

We show the efficacy of proposed spatial and joint motion and depth features in Table 1 on IRCCYN S3D video database. It is clear that from the table the estimated BGGD features alone show good performance across all distortion types and on the entire database as well. This highlights the distortion determinability of the proposed BGGD model based features. Further, it is clear that including the spatial scores shows a consistent improvement on the IRCCYN S3D video database.

4. CONCLUSION

A supervised NR VQA algorithm for natural S3D videos was proposed based on modeling the joint statistical dependencies between motion and depth subband coefficients. We showed the proposed BGGD model well captures these dependencies and estimated BGGD coefficients are distortion discriminable. The proposed VQUEMODES algorithm was evaluated on the IRCCYN and LFOVIA S3D video databases and shown the state-of-the-art performance compared to the other 2D and 3D IQA/VQA metrics. In future, we plan use the proposed model in depth scene estimation from temporal maps, visual navigation, denoising etc.

5. REFERENCES

[1] Z. P. Sazzad, S. Yamanaka, and Y. Horita, "Spatio-temporal segmentation based continuous no-reference stereoscopic video quality prediction," in *IEEE International Workshop on Quality of Multimedia Experience*, pp. 106–111, 2010.

[2] A. R. Silva, M. E. V. Melgar, and M. C. Farias, "A no-reference stereoscopic quality metric," in *Proc. SPIE*, vol. 9393, 2015.

[3] M. Solh and G. AlRegib, "A no-reference quality measure for dibr-based 3D videos," in *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2011.

[4] M. M. Hasan, J. F. Arnold, and M. R. Frater, "No-reference quality assessment of 3D videos based on human visual perception," in *IEEE International Conference on 3D Imaging*, pp. 1–6, 2014.

[5] J. Yang, H. Wang, W. Lu, B. Li, A. Badiid, and Q. Meng, "A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain," *Information Sciences*, vol. 414, pp. 133–146, 2017.

[6] K. Ha and M. Kim, "A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, p. 25252528, IEEE, 2011.

[7] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Transactions on Image Processing*, 2017.

[8] G. Jiang, S. Liu, M. Yu, F. Shao, Z. Peng, and F. Chen, "No reference stereo video quality assessment based on motion feature in tensor decomposition domain," *Journal of Visual Communication and Image Representation*, 2017.

[9] J. H. Maunsell and D. C. Van Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation," *Journal of Neurophysiology*, vol. 49, no. 5, pp. 1127–1147, 1983.

[10] 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.

[11] 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.

[12] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Image Processing, International Conference on*, vol. 3, pp. 3444–3444, IEEE Computer Society, 1995.

[13] 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.

[14] M. Jakubowski and G. Pastuszak, "Block-based motion estimation algorithmsa survey," *Opto-Electronics Review, Springer*, vol. 21, no. 1, pp. 86–102, 2013.

[15] 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.

[16] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[18] 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.

[19] B. Appina, M. K., and S. S. Channappayya, "A full reference stereoscopic video quality assessment metric," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2012–2016, March 2017.

[20] 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.

[21] 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.

[22] 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.

[23] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal, Image and Video Processing*, vol. 10, no. 4, pp. 737–744, 2016.

[24] W. Hong and L. Yu, "A spatio-temporal perceptual quality index measuring compression distortions of three-dimensional video," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 214–218, 2018.

[25] B. Appina and S. S. Channappayya, "Full-reference 3-D video quality assessment using scene component statistical dependencies," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 823–827, 2018.

[26] B. Appina, M. K., and S. S. Channappayya, "Subjective and objective study of the relation between 3D and 2D views based on depth and bitrate," in *IS&T/SPIE, Electronic Imaging*, pp. 145–150, January 2017.

[27] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[28] "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."