ALGORITHMIC ASSESSMENT OF 3D QUALITY OF EXPERIENCE FOR IMAGES AND VIDEOS

Anish Mittal, Anush K. Moorthy, Joydeep Ghosh and Alan C. Bovik

Dept. Of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, Texas - 78712.

ABSTRACT

We propose a no-reference algorithm to assess the comfort associated with viewing stereo images and videos. The proposed measure of 3D quality of experience is shown to correlate well with human perception of quality on a publicly available dataset of 3D images/videos and human subjective scores. The proposed measure extracts statistical features from disparity and disparity gradient maps as well as indicators of spatial activity from images. For videos, the measure utilizes these spatial features along with motion compensated disparity differences to predict quality of experience. To the best of our knowledge the proposed approach is the first attempt in algorithmically assessing the subjective quality of experience on a publicly available dataset.

1. INTRODUCTION

We live in an age where emerging technologies gain industry and consumer acceptance at an increasingly rapid pace. This is especially true with incubating technologies such as three-dimensional (3D) display devices. With Hollywood's increasing adoption of 3D technologies, 3D entertainment at home looks promising. Further, with BSkyB's Premiere League Soccer broadcast, the first live 3D broadcast of NFL [1], college football games shown in theaters, the 2010 Sony open golf tournament [2] and so on, 3D content is expected to make the transition from movie theaters into living rooms by next year. For example, ESPN 3D will launch its first television network in 2011 with a World Cup soccer match and expects to show at least 85 live sporting events during the first year [3]. As many experts have noted, 3D is finally here to stay [4, 5, 6].

Although there seems to be a buzz around 3D technologies, critics of the technology claim that 3D quality of experience (QoE) is unacceptable, especially during long viewing sessions. Even with advanced capture and display technologies, many viewers of 3D films have labeled them as 'unwatchable' and the discomfort associated with 3D technologies have been isolated as one of the major causes for its unpopularity in the past [7]. Given that 3D QoE is one of the most important factors in judging the palatability of visual stimuli, it is imperative that researchers in the field of quality assessment (QA) design automatic algorithms that are capable of predicting QoE of 3D stimuli. Once such automatic quality prediction is achieved, one could imagine designing algorithms that predict the optimal capture parameters given a scene content so as to maximize QoE. In particular, we are interested in predicting the optimal camera geometry in order to maximize viewing comfort, given a particular scene. A direct application of such a measure of palatability is in creation of stereoscopic content.

Although 3D QA has generated some interest in recent times [8, 9, 10, 11, 10, 12, 13, 14, 15], 3D QoE assessment algorithm design remains relatively unexplored. Traditionally, QA algorithms are classified as full-reference (FR), reduced-reference (RR) and no-reference (NR) based on the amount of information that is available to the algorithm. Since there does not exist a pristine 'reference' 3D stimulus which can be used as a baseline for comparison, 3D QoE algorithms are NR in nature. Thus, the algorithm has access to only the left and right views of a scene (and possibly the associated depth/disparity map) and needs to produce an estimate of human QoE. Our focus here is the development of such an NR 3D QoE algorithm for images and videos. Before we proceed, it is important to note that although one may continue to use the term 'quality' in the 3D realm, the term is not exactly applicable here as we do not have access to the 3 D visuosensory experience (called the cyclopean image [16]) that the human re-creates. Our algorithm has access only to the left and right views (and possibly the depth/disparity), thus making the problem of 3D QoE all the more challenging.

The rest of the paper is organized as follows: Section 2 explains details about the database we use and Section 3 describes the algorithm. In Section 4 we evaluate the proposed approach on the described publicly available database [17, 18] and demonstrate that simple statistical measures are sufficient to predict perceived quality of experience with high correlation with human perception and we conclude the paper in Section 5.

Fig. 1. Example images from the EPFL 3D image quality of experience dataset.

Fig. 2. Example frames from videos in the EPFL 3D video quality of experience dataset.

2. DATABASES USED IN THIS STUDY

The datasets that we used for evaluating stereoscopic quality of experience are those that have recently been made public by researchers at EPFL [17, 18]. There are two databases – one for images [17] and one for videos [18].

The EPFL 3D image database consists of stereoscopic images with a resolution of 1920×1080 pixels. Each scene was imaged with varying camera distances in the range $10 - 60$ cm. Note that this distance is not distance to the scene, but distance by which the camera was moved closer to the scene starting at an arbitrary reference point. The database contains 10 scenes as seen in Fig. 1. For each scene, there are 6 different camera distances to the scene being imaged. The actual database consists of 9 different scenes (since one scene was used for training human subjects) imaged at 6 depths leading to a total of 54 scenes with each being associated with a left and right view.

The EPFL 3D video database consists of videos imaged at a resolution of 1920×1080 pixels and a frame-rate of 25 fps of length 10 seconds each. Each of these scenes were again imaged at varying camera depths as described for the image database. The video database contains 6 scenes as seen in Fig. 2 imaged at 5 different camera depths. Again, the 3D video streams contain a left view and a right view.

In order to gauge human perception of quality of experience as a function of camera depth, 17 non-expert subjects participated and rated the stereoscopic images displayed on a 46" polarized stereoscopic display (Hyundai S465D) monitor at a viewing distance of 2 m on a scale of $1 - 5$ (bad, poor, fair, good and excellent) as per ITU recommendations [19]. A subjective study for 3D videos was also undertaken and consisted of 20 subjects who rated the quality of videos on the same 5-point scale.

Subjective opinion scores obtained from the above studies were averaged across subjects (after subject rejection) to produce mean opinion scores (MOS) which are representative of the perceived quality of 3D experience. Thus a total of 54 images and 30 videos with associated MOS scores are available as part of the two datasets.

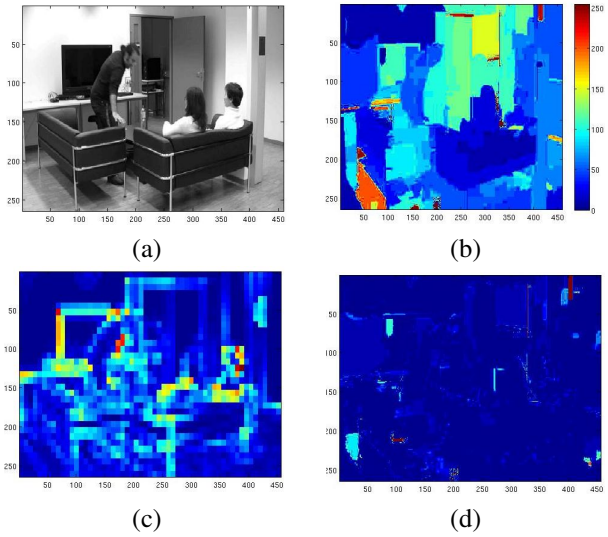

Fig. 3. (a) Left view of a video frame, (b) associated disparity map, (c) spatial activity and (d) Magnitude of motion

Our approach to no-reference (NR) QoE assessment involves extracting relevant features from these visual stimuli and regressing these features onto the MOS. In order to calibrate the regression process, we divide these datasets into various train-test combinations, train our regression module and then test how well the learned features perform in assessing the QoE.

3. ALGORITHMIC ASSESSMENT OF 3D QOE

3.1. Feature Extraction

Human stereo perception has been hypothesized to compute depth information from stereo pairs that humans receive through the two eyes [20] in order to form a cyclopean image. Hence, we first extract depth information by computing disparity between the left and right images [21]. Disparity is computed using the algorithm described in [22]. Fig. 3 (a) shows an example image and its associated disparity map is seen in Fig. 3 (b).

Thus, we now have the left and right views as well as the associated disparity maps for the stimuli in the dataset. Our hypothesis is that natural 3D images have certain statistical properties that are interpreted as 'natural' by the human observer. Deviations from this 'natural-ness' may lead to discomfort in the perception of visual stimuli thereby reducing the quality of experience. We will attempt to capture this deviation from natural-ness using simple statistical measures such as the mean, variance, skew as well as indicators of shape of

the disparity distribution. Changes in camera distance will change the statistical distributions of disparity and our hypothesis is that these changes in disparity are related to the perceived quality of experience. Apart from statistics computed from the disparity maps, we also compute spatial statistics from the left-right views in order to ensure that masking effects due to content [23] which generally influence perception are being accounted for as well.

For each 3D image (left-right pair, I_l, I_r + disparity map D) we compute the following statistical features from the disparity maps:

1. mean disparity $\mu = E[D]$,
2. median disparity $med = median(D)$,
3. disparity standard deviation $\sigma = \sqrt{E[(D-\mu)^2]}$
4. kurtosis of disparity $\kappa = E[(D-\mu)^4]/(E[(D-\mu)^2])^2$,
5. skewness of disparity $skew = E[(D-\mu)^3]/(E[(D-\mu)^2])^{(3/2)}$,
6. mean differential disparity $\mu_d = E[\delta D]$,
7. differential disparity standard deviation $\sigma_d = \sqrt{E[(\delta D - \mu_d)^2]}$
8. kurtosis of differential disparity $\kappa_d = E[(\delta D - \mu_d)^4]/(E[(\delta D - \mu_d)^2])^2$,
9. skewness of differential disparity $skew_d = E[(\delta D - \mu_d)^3]/(E[(\delta D - \mu_d)^2])^{(3/2)}$,

where the differential disparity (δD) was computed using a Laplacian operator on the disparity map. Differential disparity statistics are computed in order to capture changes in depth information [24].

To capture the nature of spatial content of the scene, we compute spatial activity for I_l and I_r from the left-right pairs. The measure of spatial activity is a modified version of the spatial indicator from [25]. Specifically, we compute the gradient of the image and estimate the variance of non-overlapping 8×8 blocks across the image. Fig. 3(c) shows an example spatial activity map for the associated image.

From the map of spatial activity S so obtained, we compute:

1. mean $\mu_s = E[S]$,
2. kurtosis $\kappa_s = E[(S-\mu_s)^4]/(E[(S-\mu_s)^2])^2$,
3. skewness $skew_s = E[(S-\mu_s)^3]/(E[(S-\mu_s)^2])^{(3/2)}$,

Such computation is undertaken for both left and right images. One could imagine pooling these measures across the two views, however, 3D perception is characterized by eye-dominance effects [26] and hence we choose to retain these individual statistics from the left-right views.

In order to evaluate the QoE of 3D videos, the above mentioned features for images are computed on a frame-by-frame basis and then averaged across frames. Other temporal pooling strategies remain interesting avenues of future research [27]. Apart from these spatial features, videos are characterized by motion information. Motion information is important for human perception and the human visual system devotes a significant amount of processing to extract motion estimates in area V5/MT of the primary visual cortex [28]. Here, we extract block motion estimates using the adaptive rood pattern search (ARPS) algorithm [29]. Block motion estimates are computed using 8×8 blocks and are a coarse approximation of the pixel-level optical flow [30]. Once block motion estimates are obtained, the difference between each 8×8 block of disparity in frame i and its motion-compensated block in frame $i-1$ is computed to form motion-compensated differences. A similar technique was applied to quality assessment of videos with success recently [31].

Once these motion compensated disparity difference maps are computed, they are pooled across each frame by computing the coefficient of variation within the frame. We note that the coefficient of variation has been used for pooling quality scores within a frame with success [27]. Finally, in order to pool these frame-level scores across the video, the median, standard deviation, kurtosis and skewness of these motion-compensated disparity differences are computed across frames and are stacked together with the computed spatial statistics. Such computation is performed twice since flows are computed on both the left and right videos separately.

Thus, for images, our feature space is 15 dimensional (5 disparity + 5 disparity gradient + 6 spatial activity), while for videos it is 25 dimensional (5 disparity + 5 disparity gradient + 6 spatial activity + 10 motion compensated disparity difference).

3.2. Feature Selection

Since the number of features computed are high compared to the size of the dataset, over-fitting is a possibility [32]. Hence, we explore two techniques - (1) principal component analysis (PCA) [33] and (2) forward feature selection (FFS) [34] for dimensionality reduction. PCA was explored initially since it provides an automated approach to dimensionality reduction, however, it is not easy to gauge significance of individual features in terms of predictive power using PCA, hence the need to explore FFS.

In PCA the feature vectors are projected onto their first n principal components where the number of principal components are chosen by cross-validation such that the features account for at least 95% of the variance; captured by 2 principal components for images and 1 principal component for videos respectively (averaged across trials). In FFS, we first choose that feature that correlates the best with subjective scores on

Method	Mean	Standard deviation
PCA	0.79	0.08
FFS	0.86	0.11

Table 1. 3D Image QoE: Spearman's Rank Ordered Correlation Coefficient (SROCC) values across ${}^9C_6$ train-validate-test trials.

Method	Mean	Standard deviation
PCA	0.76	0.25
FFS	0.68	0.28

Table 2. 3D video QoE: SROCC values across ${}^6C_4$ train-validate-test trials.

the training set; then, that feature which correlates the best with subjective data *in conjunction with* the first feature is chosen and so on [34]. This process continues until a stopping criterion is reached. In our implementation this criterion is decided through cross-validation. Specifically, the final set of features picked are those that are selected the most across validation trials. The number of selected features is the median number of features selected across validation trials.

Thus, for both PCA and FFS, at the end of the training stage, a set of features that predict the training/validation data the best are chosen. In the case of PCA, test features are projected on to the space formed by the principal components from the training set. For FFS, features selected from the test set are the same as the ones obtained from FFS on the training set.

In order to map the features onto subjective scores we used a simple linear regression model for performance evaluation:

$$y = \vec{x}^T \vec{\gamma} + \eta \tag{1}$$

where $\vec{x}$ is the vector of features, $\vec{\gamma}$ is the weight vector whose parameters need to be estimated and η is a constant which needs to be estimated as well.

4. PERFORMANCE EVALUATION

4.1. 3D image quality of experience

Recall that the image dataset consists of 9 scenes imaged at 6 camera distances. We utilize $4(\times 6)$ images for training, $2(\times 6)$ images for validation and $3(\times 6)$ images for testing. Since we wish to demonstrate that the proposed approach is robust across contents, we used all possible combinations of the dataset - ${}^9C_6$ - to form the above mentioned training and validation/test sets. Results reported in Table 1 are the mean and standard deviation of Spearman's rank ordered correlation coefficient (SROCC) between the features regressed as described above and the subjective opinion score across these combinations on the test set for PCA and FFS. SROCC of 1 indicates a perfect correlation. It should be clear that the proposed approach performs well in terms of correlation with human perception.

Finally, our motivation for the use of FFS was that one could intuit on which features are more relevant in predicting quality of experience. Hence in Fig.4 we plot a histogram of features selected across the ${}^9C_6$ train and validate/test trials using FFS. One would conjecture that the mean and median of disparity maps are of tremendous importance in gauging quality of experience, however, they alone are not sufficient.

4.2. 3D video quality of experience

From the database of 6 videos imaged at 5 distances, we use $3(\times 5)$ videos for training, $1(\times 5)$ videos for validation (i.e., leave out one validation) and $2(\times 5)$ videos for testing. Again, to ensure that all contents are evaluated, all possible combinations - ${}^6C_4$ - are used to produce train and validate/test samples. Mean and standard deviation of SROCC values across these ${}^6C_4$ combinations are reported in Table 2 for PCA and FFS. Again, the statistical features seem to perform well in terms of correlation with human perception across contents.

Fig.5 shows a histogram of features chosen across these permutations for FFS. Again, the results are intuitive. Motion-compensated disparity plays an important part in assessing the quality of experience of 3D video. This result agrees with psychovisual evidence of the importance of motion information [28] as well as evidence from 2D video quality assessment [31].

It is interesting to note that even though both left and right flow features are selected, only one of them has a significant magnitude at a time (as observed by us), since the disparity map already incorporates masking effects and flow compensation need not be done twice.

5. CONCLUSION AND FUTURE WORK

We proposed a no-reference objective quality of experience assessment model to evaluate the perceived quality of 3D experience for images and videos. We evaluated the proposed approach on publicly available datasets for 3D quality of experience and demonstrated that the proposed algorithm correlates well with human perception of quality. We observed that the comfort associated with viewing stereoscopic stimuli reduces with increasing distances to the reference point and that this effect is more pronounced in indoor scenes with higher disparity gradients.

Future work would involve utilizing the proposed model to predict the optimal depth distribution for specific scene contents. One could imagine the proposed model being applied to gauge appropriate camera distances for capturing 3D scenes using a parallel baseline setup. Also, exploring other feature selection mechanisms such as step-wise variable selection would be an interesting proposition.

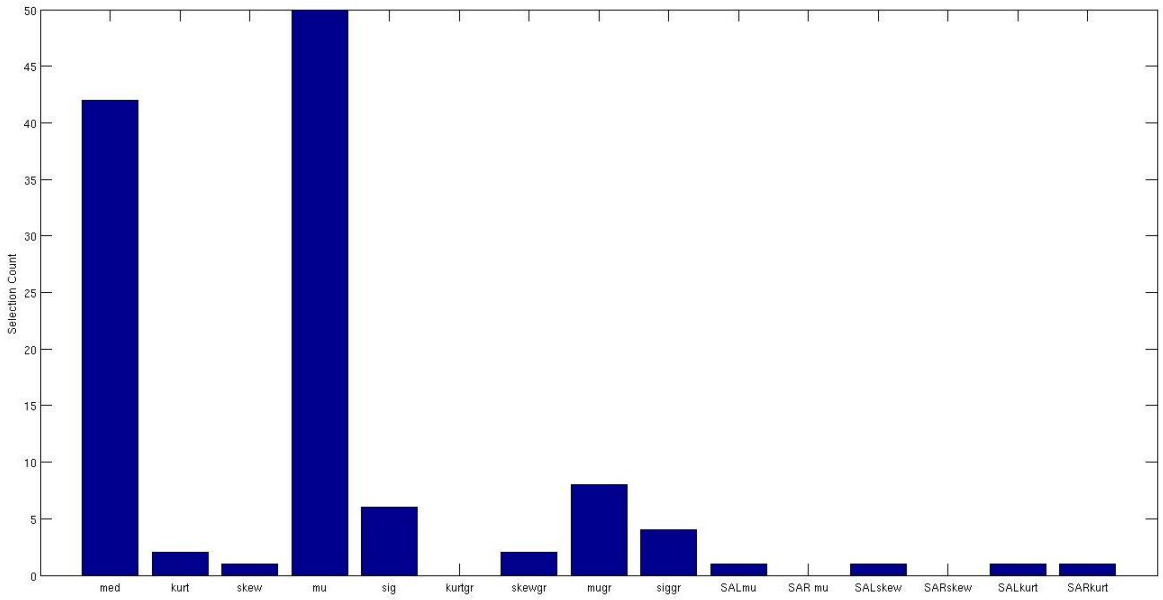

Fig. 4. 3D Image QoE: Histogram of features chosen across ${}^{9}C_{6}$ train and validate/test $(4-2-3)$ trials using FFS. med, kurt, skew, mu, sig = median, kurtosis, skewness, mean and standard deviation from disparity maps; kurtgr, skewgr, mugr, siggr = mean, kurtosis, skewness and standard deviation from the differential disparity map. SAL = spatial activity from left image, SAR = spatial activity from right image, SARmu/SALmu = mean spatial activity, SARskew/SALskew = skewness of spatial activity, SALkurt/SARkurt = kurtosis of spatial activity.

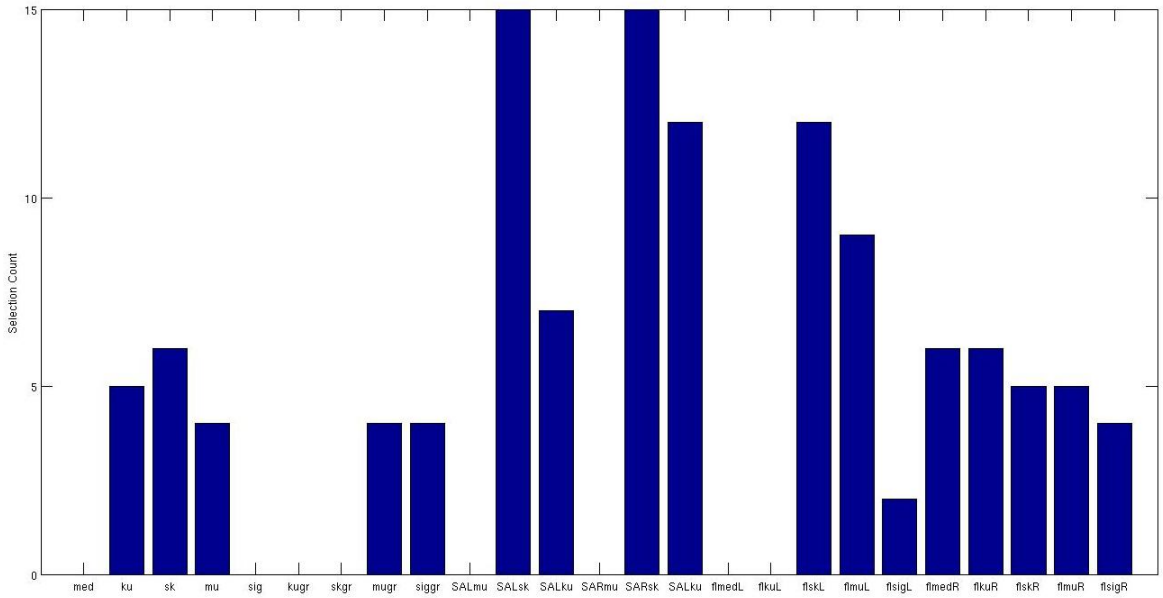

Fig. 5. 3D video QoE: Histogram of features chosen across ${}^{6}C_{4}$ train and validate/test$(3-1-2)$ trials using FFS.med, kurt, skew, mu, sig = median, kurtosis, skewness, mean and standard deviation from disparity maps; kurtgr, skewgr, mugr, siggr = mean, kurtosis, skewness and standard deviation from the differential disparity map. SAL = spatial activity from left image, SAR = spatial activity from right image, SARmu/SALmu = mean spatial activity, SARskew/SALskew = skewness of spatial activity, SALkurt/SARkurt = kurtosis of spatial activity. The prefix 'fl' refers to motion compensated disparity features and the suffixes L and R represent the right and left streams.

6. REFERENCES

[1] PR NewsWire, "3ality digital's live 3d broadcast of an nfl game," http://www.prnewswire.com/news-releases/3ality-digitals-first-ever-live-3d-broadcast-of-an- nfl-game-named-one-of-sports-illustrateds-innovations-of-the-decade-80318427.html.

[2] Bloomberg BusinessWeek, "Golf masters tournament to be broadcast in 3d," http://www.businessweek.com/news/2010-03-16/golf-s-masters-tournament- to-be-broadcast-in-3-d-tv-correct-.html.

[3] ESPN News, "Espn 3d," http://sports.espn.go.com/espn/news/story?id=4796555.

[4] L.M.J. Meesters, W.A. IJsselsteijn, and P.J.H. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, 2004.

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

[6] X. Wang, M. Yu, Y. Yang, and G. Jiang, "Research on subjective stereoscopic image quality assessment," vol. 7255, pp. 725509, 2009.

[7] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: a review," *Journal of Imaging Science and Technology*, vol. 53, pp. 030201, 2009.

[8] B. Alexandre, L.C. Patrick, C. Patrizio, and C. Romain, "Quality Assessment of Stereoscopic Images," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2009.

[9] X. Wang, M. Yu, Y. Yang, and G. Jiang, "Research on subjective stereoscopic image quality assessment," in *Proceedings of SPIE*, 2009, vol. 7255, p. 725509.

[10] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation," *ACM Transactions on Applied Perception (TAP)*, vol. 3, no. 2, pp. 109, 2006.

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

[12] Sazzad Z. M. P., Yamanaka S., Kawayoke Y., and Horita Y., "Stereoscopic image quality prediction," *Proceedings of IEEE QoMEX, San Diego, CA, USA*, pp. 180–185, 2009.

[13] C. 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, no. 16, pp. 963–965, 2008.

[14] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," *Proceedings of SPIE Stereoscopic Displays and Applications XIX*, vol. 6803, pp. 680305, 2008.

[15] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proceedings of 15th European Signal Processing Conference (EUSIPCO)*, 2007.

[16] B. Julesz, T.V. Papathomas, and F. Phillips, *Foundations of cyclopean perception*, University of Chicago Press Chicago:, 1971.

[17] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortions on the quality of stereoscopic images," *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2010.

[18] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Electronic Imaging (EI), 3D Image Processing (3DIP) and Applications*, 2010.

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

[20] I.P. Howard and B.J. Rogers, *Seeing in depth*, I. Porteous, 2002.

[21] E. Trucco and A. Verri, *Introductory techniques for 3-D computer vision*, Prentice Hall New Jersey, 1998.

[22] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Pattern Recognition, 18th International Conference on*, 2006, vol. 3.

[23] Y. Yang and R. Blake, "Spatial frequency tuning of human stereopsis," *Vision research*, vol. 31, no. 7-8, pp. 1176–1189, 1991.

[24] Y. Liu, L.K. Cormack, and A.C. Bovik, "Natural scene statistics at stereo fixations," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 161–164.

[25] MH Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[26] N.K. Logothetis and J.D. Schall, "Binocular motion rivalry in macaque monkeys: eye dominance and tracking eye movements," *Vision research*, vol. 30, no. 10, pp. 1409–1419, 1990.

[27] K. Seshadrinathan and A.C. Bovik, "Motion-based perceptual quality assessment of video," *Proc. SPIE-Human Vision and Electronic Imaging*, 2009.

[28] S.E. Palmer, *Vision science: Photons to phenomenology*, MIT press Cambridge, MA., 1999.

[29] Y. Nie and K.K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transactions on Image Processing*, vol. 11, no. 12, pp. 1442–1449, 2002.

[30] D.J. Fleet and A.D. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 1990.

[31] A. K. Moorthy and A. C. Bovik, "A motion compensated approach to video quality assessment," *Asilomar Conference on Signals, Systems and Computers*, 2009.

[32] C.M. Bishop et al., *Pattern recognition and machine learning*, Springer New York:, 2006.

[33] IT Jolliffe, *Principal component analysis*, Springer verlag, 2002.

[34] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.