EVQA: AN ENSEMBLE-LEARNING-BASED VIDEO QUALITY ASSESSMENT INDEX

Joe Yuchieh Lin [*], *Chi-Hao Wu* [*], *Ioannis Katsavounidis* [†], *Zhi Li* [†], *Anne Aaron* [†], *and C.-C. Jay Kuo* [*]

[*] University of Southern California, Los Angeles, California, USA
[†] Netflix, Los Gatos, California, USA

ABSTRACT

A full-reference video quality assessment (VQA) method, called the ensemble-learning-based video quality assessment (EVQA) index, is proposed in this work. As compared with previous learning-based VQA methods, it has two unique features. First, EVQA adopts a frame-based learning mechanism to address the limited training data problem. Second, a dynamic image quality assessment(IQA) fusion scheme is developed by taking three factors into account: the spatial complexity and temporal context of a frame in a video source and the strength of IQA indices. In the test stage, EVQA applies the derived IQA fusion rule to different frames and take an average of the frame-based scores to generate the final video quality score. The superior performance of the proposed EVQA index is demonstrated by experimental results conducted on both LIVE and MCL-V video databases.

1. INTRODUCTION

Video streaming services are growing rapidly nowadays, and the cloud platform is adopted for video streaming on a massive scale. Automatic video quality assessment (VQA) plays an important role to measure viewer's experience in such an application. Although the mean squared error (MSE) is commonly used for video quality measure, it is not well correlated to human visual experience [1]. Thus, there has been active research on developing better objective VQA indices to meet this urgent need.

Generally speaking, one can categorize VQA methods into two types: formula-based and learning-based VQA methods. The development of formula-based VQA indices are hindered by two major factors. First, video contents are very diversified, and it is difficult to model them mathematically. Second, the human visual system (HVS) is too complex to be fully understood although there are scattered results on psychovisual theory. Due to the challenge in source (video) and receiver (human) modeling, the performance of today's formula-based VQA indices is still far from perfection. The development of learning-based VQA indices is constrained by the limited amount of labeled data. Although there are quite a few image quality assessment (IQA) databases available to the public, the number of VQA databases is very limited since it is expensive to conduct a large scale subjective test on video.

A full-reference video quality assessment (VQA) method, called the ensemble-learning-based video quality assessment (EVQA) index, is proposed in this work. As compared with previous learning-based VQA methods, it has two unique features. First, EVQA adopts a frame-based learning mechanism to address the problem of a limited amount of training data. Second, a dynamic IQA fusion scheme is developed by taking three factors into account: the spatial complexity and temporal context of a frame in a video source and the strength of IQA indices. In the test stage, EVQA applies the derived IQA fusion rule to different frames and take an average of the frame-based scores to generate the final video quality score. The main contribution of this work lies in an enhanced learning procedure by partitioning frames into multiple groups based on their contents and contexts as detailed in Section 3. The superior performance of the proposed EVQA index is demonstrated by experimental results conducted on both LIVE and MCL-V video databases.

The rest of this paper is organized as follows. The background of our research is reviewed in Section 2. The proposed EVQA index is presented in Section 3. Experimental results are shown in Section 4. Finally, concluding remarks are given in Section 5.

2. BACKGROUND REVIEW

There are two approaches to the design of an IQA/VQA index; namely, formula-based and learning-based. They are reviewed below. In the formula-based approach, a closed form mathematical model is derived to predict perceptual quality, such as frame-based structural similarity (SSIM) [2], visual information fidelity (VIF) [3], feature similarity (FSIM) [4], and video additive impairments and detail losses measure (VADM) [5]. However, it is extremely difficult to provide a good mathematical HVS model to cover a wide range of video collections. To give an example, Li *et al.* [5] used Daly's contrast sensitivity function (CSF) [6] to improve the performance of VQA indices. However, there are more visual properties than contrast in the HVS, including luminance adaptation, visual saliency among others, the applicability of which is rather limited.

In learning-based approaches, a statistical model is built to model the relation between features of training image/video data and their mean opinion scores (MOS). Then, it is used to predict the quality of unseen test video. This approach has been used by researchers to design IQA indices in recent years. Narwaria and Lin [7] extracts the structural information in images with the singular value decomposition and then use the support vector regression (SVR) to map the feature to MOS. Liu *et al.* [8] proposed a multi-method fusion (MMF) IQA index, where a regression approach is used to combine scores of multiple IQA indices. The MMF score is obtained by a non-linear fusion of scores computed by multiple methods with suitable weights obtained by a training process. The MMF index offers the state-of-the-art IQA results in several popular databases, including LIVE [9] and TID2008 [10].

The learning-based approaches have also been applied to the design of VQA indices. The fusion-based VQA (FVQA) technique was proposed in [11]. The FVQA method treats each video clip as a single data sample. In the training stage, it first classifies video clips based on their spatial and temporal complexities into several groups to reduce intra-group content diversity. Then, it learns the fusion rule of multiple VQA indices in each group. In the testing stage, the FVQA index first classifies a test video clip into a group and then applies the fusion rule in that group for VQA score prediction.

Since the development of formula-based VQA indices is hindered by video content diversity and HVS complexity, we adopt a learning-based approach in this work. It is however important to emphasize that the number of training images in image quality databases [9, 10] is significantly larger than that in video quality databases. The accuracy of a statistical VQA model can be severely affected by the small size of the training data. This is the main issue to be addressed in our current work. Two exemplary video quality databases are used in our experiments. They are the LIVE database [12] and the MCL-V database [13]. The LIVE database contains 80 video clips of resolution 768×432 and with a duration of 10 seconds. They were coded by H.264 and MPEG-2. The MCL-V database contains 12 source video clips of resolution 1080p and with a duration of 6 seconds. There are 96 distorted video clips due to compression and scaling.

3. PROPOSED EVQA INDEX

Motivation and Overview. A video stream is composed by image frames, where frame-to-frame variation is usually small except for scene change. It is a commonly believed fact that perceptual quality remains stable within a short period of time. This property was exploited in [5, 14], where a spatial (or frame-level) quality index is first computed for each frame independently and the index scores across multiple consecutive frames can be weighted by a temporal pooling method. In this way, a VQA index can be constructed from frame-level IQA indices. To tackle the problem of limited training data, our proposed method adopts a frame-based learning mechanism, inspired by the same principle. Each source video clip in the MCL-V database lasts for 6 seconds and the frame rate is 30 frames per second (fps). The total number of frames for one sequence is 180 frames. In other words, each training video clip can offer 180 data samples, instead of just one. There is however a missing link in the aforementioned strategy; namely, the frame-level MOS score is not available during the training process. Since all video quality databases contain short and homogeneous video clips, it is assumed that the MOS of the whole sequence can be used as an approximate ground truth of its frames. This assumption will be verified in Section 4.

The EVQA method consists of three steps in the training phase. Step 1: Feature Extraction. Several IQA methods are applied to each individual frame and their scores are stored as its feature vector. The raw scores of IQA indices are properly normalized to match the MOS value. Step 2: Frame Space Partitioning. The frame space is partitioned into several subspaces to enhance the learning performance. Step 3: IQA index Fusion. The fusion rule of combining multiple IQA indices into one single IQA score for a frame is learned in each partitioned frame subspace.

In the testing process, the EVQA method predicts the quality index of each frame by following the above steps. After that, all predicted frame scores are integrated to generate one MOS value for a short test video clip via *temporal pooling*. Since feature extraction is straightforward, we will elaborate on the following three topics below: frame space partitioning, IQA index fusion and temporal pooling.

Frame Space Partitioning. The main purpose of frame space partitioning is to allow more efficient learning rule in a smaller subspace, where frames share properties of higher similarity. This can be done based on spatial, temporal and quality/distortion properties of frames. The spatial and temporal complexities are related to the spatial and temporal masking effects of HVS. For the quality/distortion property, the predicted performance of an IQA index can be exploited. That is, each IQA index has its own strength in assessing some distortion types [8] and, if two frames can be well predicted by a common set of IQA indices, they must share certain similarity in their quality/distortion property.

Spatial and temporal complexities are computed based on the undistorted reference frames. The spatial information (SI) and the temporal information (TI) introduced in [15] are two well-known parameters for video sequences. However, they are not suitable for our purpose since we are concerned with the properties of a single frame. Some modifications are needed, and we call extended SI (ESI) and extended TI (ETI) the modified metrics.

For the ESI, we first obtain the edge magnitude map G_M of frame F_n using the 3×3 Sobel filter. Then, the ESI for this

frame is defined as

$$ESI_n = \frac{std[G_M(F_n)]}{mean[G_M(F_n)]}\,. \tag{1}$$

Complex scenes with a large amount of texture have a larger ESI value. For the ETI, the basic idea is to compute the pixel-based luminance difference of two adjacent frames. Sequences with large motion have large ETI values. Since the fine structure of the frame data, such as film noise, will have a negative impact on the accuracy of ETI, we apply the 5×5 Gaussian filter to their pixel difference, which is equivalent to taking the difference after we filter out each frame by the same Gaussian filter. Then, the ETI is defined as

$$ETI_n = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} D_n(x,y)\,. \tag{2}$$

where W and H are the width and the height, correspondingly of the nth frame and

$$D_n(x,y) = |G * (F_n(x,y) - F_{n-1}(x,y))| \tag{3}$$

is the absolute value of the Gaussian-smoothed frame difference, and where G is the Gaussian filter.

Besides spatial and temporal complexities, it is desired to classify image frames based on their distortion type. However, it is difficult to obtain this information directly, yet it is possible to be obtained indirectly by analyzing its IQA scores. This analysis is conducted with respect to frames in the training set. Suppose that there are T training frames. For a given IQA index, we can divide all training frames into C clusters of equal size $N = T/C$, based on its score distribution. We map raw IQA scores in one cluster, denoted by x, to a normalized score, Q, using a logistic function [4] as follows:

$$Q = \beta_1 \cdot (0.5 - \frac{1}{1 + e^{\beta_2(x-\beta_3)}}) + \beta_4 \cdot x + \beta_5, \tag{4}$$

where $\beta_i, i = 1...5$, are the fitting parameters determined by known IQA/MOS score pairs. Eq. (4) is used to convert an IQA score of an arbitrary range to a suitable range which is compatible with measured MOS values. After the score conversion, we can compute the root-mean-squared-error (RMSE), denoted by E, between Q and MOS in that cluster via

$$E(Q, MOS) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Q_n - MOS_n)^2}, \tag{5}$$

where n is a frame index, MOS_n is its MOS value and Q_n is its transformed IQA index value. Furthermore, we can choose a threshold value to determine if an IQA method performs well in a cluster. For example, the RMSE values of 8 clusters are shown in Fig. 1. By setting the threshold value to $E = 1$,

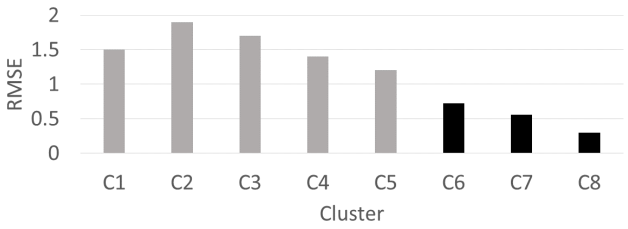
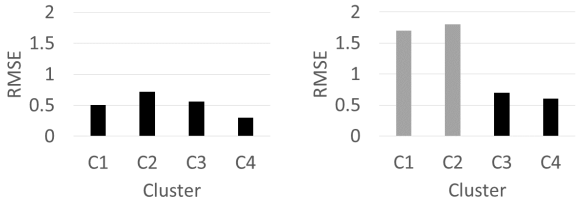


Fig. 1: The plot of the RMSE of the predicted MOS values against the actual ones in 8 clusters for an exemplary IQA index, where a black and a gray bars indicate that its RMSE is lower and higher than a pre-selected threshold value, respectively.

we see that the IQA index performs well for frames in Cluster Nos. 6-8 but poorly for frames in Cluster Nos. 1-5.

If an IQA index performs equally well (or poorly) for all clusters as shown in Fig. 2 (a), it cannot be used to partition the frame space. On the other hand, if it performs well for some clusters but poorly for other clusters as shown in Fig. 2 (b), we can use it to partition the frame space into two subspaces according to its preference - favored and unfavored subspaces.



(a) Index without preference (b) Index with preference

Fig. 2: Illustration of an IQA index (a) without and (b) with preference.

Furthermore, we can use a sequence of IQA indices with preference to partition a frame space into multiple subspaces as illustrated in Fig. 3, where each split is defined by one IQA index. In this figure, the favored and unfavored subspaces of the first IQA index is denoted by A and A^c, respectively. Similarly, the frame space can be partitioned by another IQA index into the favored and unfavored subspaces denoted by B and B^c, respectively. Then, the frame space can be decomposed into four subspaces as shown in the third stage of Fig. 3.

The frame space partitioning process can be organized as a binary tree as shown in Fig. 4. Each partition creates two children, and grows the tree to the next level. The partition should stop if the number of frames in a node is too small since each group should have a sufficient number frames for the learning purpose. On the other hand, for nodes that have

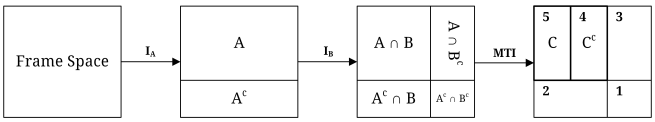


Fig. 3: Frame space partitioning using multiple IQA indices with preference.

a large number of frames, after we exhaust all IQA indices, we can use the frame′s ETI and ESI value to further partition them. Then, we can use frame's ETI and ESI to partition them.

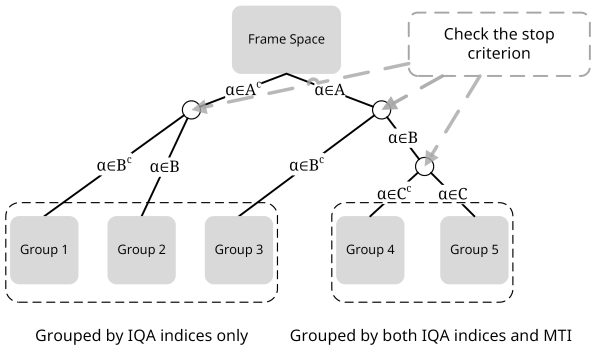


Fig. 4: Illustration of frame space partitioning using a binary tree structure, where the stop criterion is checked at each node.

IQA index Fusion. In our experiment, six state-of-the-art IQA indices [16, 2, 3, 4, 17, 18] are included in the IQA candidate pool. For each partitioned frame subspace, we use the sequential forward method selection (SFMS) scheme [8, 11] to select a set of IQA indices to fuse so as to optimize an objective function such as the Pearson linear correction coefficient (PCC) value. The SFMS scheme is a greedy search algorithm that selects the optimal IQA index in the candidate pool to yield a better prediction at each iteration. The iteration will terminate if the improvement becomes negligible. The SFMS scheme is determined by training data while the fusion rule is also learned through SVR from training data in each frame subspace.

Temporal Pooling. Temporal pooling is necessary to generate the final MOS value for the entire test video based on the predicted MOS value of each individual frame. Several temporal pooling methods such as the mean, median, Minkowski, percentile was studied and compared in [19, 20]. There is however no universal method that offers the best performance for all video contents. We adopt a simple average scheme here, which is justified by experimental results in Section 4.

4. EXPERIMENTAL RESULTS

We present experimental results in two parts in this section. In the first part, we study the relationship between the frame-level and the sequence-level quality indices to justify two items: 1) the assumption that the sequence-level MOS can be used to approximate the frame-level MOS, and 2) the adoption of simple averaging as the temporal pooling method in EVQA.

Relationship between Frame-Level and Sequence-Level MOS Values. The frame-to-frame quality level is assumed to be stable for a short period if no scene change occurs. To verify this assumption, we plot the predicted frame-level MOS as a function of the frame index for the BC (Birds in Cage) sequence coded under "good" quality in the MCL-V quality database in Fig. 5. We see that the predicted frame-level MOS is nearly constant.

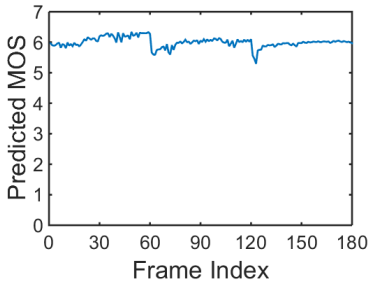


Fig. 5: The predicted frame-level MOS value is plotted as a function of the frame index for the BC sequence coded under "good" quality, where predicted sequence-level MOS is 6.81 by simple averaging while the true MOS value is 7.06.

The MCL-V video quality database consists of 12 sequences with five quality levels caused by different coding bitrates. We show the mean (μ) and the standard deviation (σ) of the predictive frame-level MOS values for all of them, in four quality levels (good, fair, poor and bad) in Table 1. The first column is the acronym for the title of each sequence. When the standard deviation value is low, it means that the frame-level MOS is nearly a constant. There are several sequences with larger standard deviation values such as DK (Dance Kiss), EA (El Fuente A), EB (El Fuente B), FB (Fox Bird) and TN (Tennis). These sequences were shot with more complex camera-object relative motion; thus they deviate slightly from the homogeneous frame-level MOS assumption.

To examine such a deviation in detail, we plot the predicted frame-level MOS as a function of the frame index for the DK sequence coded under "good" quality in the MCL-V quality database in Fig. 6. We do observe the fluctuation of the predicted frame-level MOS values between frames 45-90 caused by camera-object relative motion. However, the predicted sequence-level MOS (6.48) is still close to its ground

Table 1: Mean and standard deviation of frame scores with compression distortion in MCL-V

Qual. Level	Good		Fair		Poor		Bad	
Seq. Title	μ	σ	μ	σ	μ	σ	μ	σ
BB	6.17	0.43	4.92	0.58	2.97	0.67	1.82	1.53
BC	6.81	0.09	5.79	0.59	2.69	0.59	0.71	0.26
BQ	6.88	0.19	6.42	0.52	3.58	0.55	1.80	0.64
CR	6.76	0.50	4.50	0.41	3.22	0.42	1.29	0.64
DK	6.48	1.05	5.42	1.88	2.74	1.65	0.42	1.72
EA	6.64	0.82	4.37	0.92	1.72	0.70	0.29	0.06
EB	4.92	0.92	3.48	1.18	1.52	0.68	1.76	0.79
KM	6.24	0.50	5.00	0.90	2.81	0.71	1.05	0.52
FB	6.00	0.72	3.57	1.08	1.37	0.89	1.43	1.26
OT	6.55	0.24	6.34	0.19	2.80	0.49	0.73	0.28
SK	6.75	0.08	5.34	0.52	3.35	0.42	0.75	0.36
TN	6.32	0.47	4.82	0.58	4.12	0.69	1.29	0.84

truth (6.63).

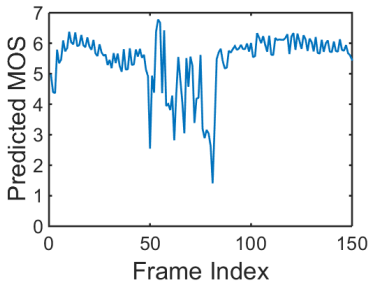


Fig. 6: The predicted frame-level MOS value is plotted as a function of the frame index for the DK sequence coded under "good" quality, where predicted sequence-level MOS is 6.48 by simple averaging while the true MOS value is 6.63.

Performance Comparison of VQA Indices. The performance of the proposed EVQA method is evaluated on the MCL-V database [13] and the coding distortion of the LIVE video database [12]. We follow the validation process proposed by VQEG in [21]. First, IQA index scores [16, 2, 3, 4, 22, 17, 23, 18] are mapped by the logistic function given in Eq. 4. Then, we consider three commonly used performance measures: 1) the Pearson correlation coefficient (PCC), 2) the Spearman rank-order correlation coefficient (SROCC), and 3) the root mean squared error (RMSE). PCC computes the correlation between the true and predicted MOS values. SROCC measures prediction monotonicity. RMSE calculates the error between the true and predicted MOS values.

We adopt the 10-fold cross-validation strategy to select training and testing sets in the experiments. The performance of EVQA is compared with several benchmarking IQA and VQA indices. If an IQA index is used, its simple averaging is adopted to yield the final sequence-level MOS value. PCC, SROCC and RMSE results against the LIVE and the MCL-V databases are shown in Tables 2 and 3, respectively. Clearly, EVQA outperforms all other indices in every performance measure in both databases.

Table 2: Performance comparison of video quality indices for video clips in the LIVE video quality database with compression distortion (H.264 and MPEG-2).

	PCC	SROCC	RMSE
PSNR	0.478	0.449	9.034
VIF [3]	0.600	0.607	8.236
MSSIM [17]	0.591	0.692	8.294
FSIM [4]	0.634	0.698	7.955
GSM [18]	0.614	0.658	8.117
ST-MAD [14]	0.838	0.825	5.607
VADM [5]	0.847	0.850	5.469
EVQA	0.934	0.926	3.664

Table 3: Performance comparison of video quality indices for video clips in the MCL-V video quality database.

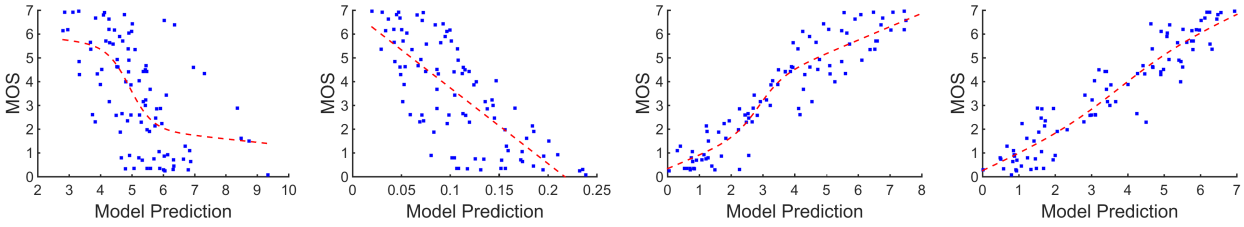
	PCC	SROCC	RMSE
PSNR	0.476	0.426	1.984
VIF [3]	0.660	0.655	1.666
MSSIM [17]	0.621	0.623	1.740
FSIM [4]	0.755	0.747	1.455
GSM [18]	0.709	0.711	1.565
ST-MAD [14]	0.634	0.623	1.714
VADM [5]	0.742	0.752	1.489
FVQA [11]	0.945	0.932	0.727
EVQA	0.956	0.947	0.652

Fig. 7 shows the scatter plots of four leading methods in Table 3, where each dot gives the predicted MOS value and the actual MOS value in its x-coordinate and y-coordinate, respectively, for each test sequence in the MCL-V database. The red dash line indicates the optimal regression curve for these points. The ideal case is a straight line starting from zero along either the positive or the negative 45-degree direction with little deviation. The ST-MAD [14] regression curve is not straight while its data points are too spread out. The VADM [5] has a more straight regression line, yet its data points are still quite spread out. In contrast, data points in FVQA [11] and EVQA are much closer to their regression lines. Furthermore, the regression line of EVQA is more straight than that of FVQA.

5. CONCLUSION AND FUTURE WORK

A new VQA index called EVQA was proposed to assess the quality of streaming video. Under the quasi-stationary assumption, the MOS value of a short video clip can be approximated by the running average of the corresponding MOS of each frame in that clip. In this way, we convert the sequence-level VQA problem to the frame-level IQA problem. The frame-level IQA problem fits the learning framework better due to the existence of a larger amount of training samples. The frame space partitioning and IQA index fusion techniques were adopted to enhance the performance of frame-level quality prediction.

The length of a streaming video program can be quite



(a) STMAD [14] (b) VADM [5] (c) FVQA [11] (d) EVQA

Fig. 7: Scatter plots and their regression curves for all sequences in the MCL-V database using (a) ST-MAD, (b) VADM, (c) FVQA and (d) EVQA indices.

long, and it will contain various scenes with scene changes in between. In practice, we have to divide one long video sequence into homogeneous segments and perform VQA on each segment. Thus, the final VQA index should be a function of time. The segmentation of a long video program into proper units that have a constant VQA value is under our current investigation.

6. REFERENCES

[1] Bernd Girod, "What's wrong with mean-squared error?," in *Digital images and human vision*. MIT press, 1993, pp. 207–220.

[2] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[3] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[4] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *Image Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2378–2386, 2011.

[5] Songnan Li, Lin Ma, and King Ngi Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 7, pp. 1100–1112, 2012.

[6] Scott J Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 2–15.

[7] Manish Narwaria and Weisi Lin, "Objective image quality assessment based on support vector regression," *Neural Networks, IEEE Transactions on*, vol. 21, no. 3, pp. 515–519, 2010.

[8] Tsung-Jung Liu, Weisi Lin, and C.-C. Jay Kuo, "Image quality assessment using multi-method fusion," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1793–1807, 2013.

[9] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *Image Processing, IEEE Transactions on*, vol. 15, no. 11, pp. 3440–3451, 2006.

[10] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, M. Carli, and F. Battisti, "TID2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[11] Joe Yuchieh Lin Lin, Tsung-Jung Liu, Eddy Chi-Hao Wu, and C.-C. Jay Kuo, "A Fusion-based Video Quality Assessment (FVQA) Index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014.

[12] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1427–1441, 2010.

[13] Joe Yuchieh Lin, Rui Song, Chi-Hao Wu, TsungJung Liu, Haiqiang Wang, and C.-C. Jay Kuo, "MCL-V: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, pp. –, 2015, (to appear in print).

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

[15] ITU, "Recommendation ITU-T P.910, Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union, Geneva, Switzerland*, vol. 910, 1999.

[16] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 935–949, 2011.

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

[18] Anmin Liu, Weisi Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1500–1512, April 2012.

[19] Snježana Rimac-Drlje, M Vranjes, and Drago Zagar, "Influence of temporal pooling method on the objective video quality evaluation," in *Broadband Multimedia Systems and Broadcasting, 2009. BMSB'09. IEEE International Symposium on*. IEEE, 2009, pp. 1–5.

[20] Michael Seufert, Martin Slanina, Sebastian Egger, and Meik Kottkamp, "to pool or not to pool: A comparison of temporal pooling methods for http adaptive video streaming," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 2013, pp. 52–57.

[21] Video Quality Experts Group (VQEG), "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II," 2003.

[22] Theophano Mitsa and Krishna Lata Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*. IEEE, 1993, vol. 5, pp. 301–304.

[23] Wufeng Xue, Lei Zhang, Xuanqin Mou, and AC. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 684–695, Feb 2014.