A Fusion Approach to Video Quality Assessment Based on Temporal Decomposition

Tsung-Jung Liu*, Weisi Lin† and C.-C. Jay Kuo*
*Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA
E-mail: liut@usc.edu, cckuo@sipi.usc.edu Tel: +1-213-7404658
† School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore
E-mail: wslin@ntu.edu.sg Tel: +65-67906651

Abstract— **In this work, we decompose an input video clip into multiple smaller intervals, measure the quality of each interval separately, and apply a fusion approach to integrating these scores into a final one. To give more details, an input video clip is first decomposed into smaller units along the temporal domain, called the temporal decomposition units (TDUs). Next, for each TDU that consists of a small number of frames, we adopt a proper video quality metric (specifically, the MOVIE index in this work) to compute the quality scores of all frames and, based on the sociological findings, choose the worst scores of TDUs for data fusion. Finally, a regression approach is used to fuse selected worst scores from all TDUs to get the ultimate quality score of the input video as a whole. We conduct extensive experiments on the LIVE video database, and show that the proposed approach indeed improves MOVIE and is also competitive with other state-of–the-art video quality metrics.**

I. Introduction

In recent years, photo and video sharing on the Internet becomes much more popular and available than before because of the development of social networks and digital mobile devices. People can share or watch videos on some specific websites, such as Facebook or YouTube. In addition, video conferences are also often used to replace face-to-face meeting these days. Even more video applications and services are expected in the near future, as a result of advanced video coding and communications. Therefore, how to assess visual quality and assure acceptable quality of experience (QoE) for digital images and videos in an objective manner become an increasingly critical and interesting topic in the research community.

One obvious way to implement video quality assessment (VQA) is to apply a still image quality assessment metric (e.g., SSIM [1]) on a frame-by-frame basis. The quality of each frame is evaluated independently, and the global quality of the video sequence can be obtained by a simple time-average of quality scores from each frame. However, the performance of such methods is not satisfactory, since the important temporal characteristics of visual signals are not sufficiently accounted for.

It is believed that considering temporal information along with spatial domain would improve the quality prediction performance [2]-[4]. For example, Wang *et al.* [5] applied SSIM to video quality assessment by employing the motion analysis to give different weightings to the level quality scores. This approach is often known as V-SSIM, which has been demonstrated to perform better than other methods as reported in the Video Quality Experts Group (VQEG) Phase I final report [6]. Speed-SSIM [7] is another method that incorporates a model of human visual speed perception [8] by considering the visual perception process in an information communication framework. Consistent improvement over existing VQA algorithms is observed in the validation with the VQEG Phase I test data set [9].

Moreover, National Telecommunications and Information Administration (NTIA) developed a VQA metric, called Video Quality Metric (VQM) [10]. Due to the excellent performance in the VQEG Phase II validation tests, VQM has been adopted as a national standard by the American National Standards Institute (ANSI) and also as International Telecommunications Union Recommendations [11, 12].

Although VQM has been introduced as the national standard for all applications, some efforts have been further made to improve the performance of video quality metrics. The latest, most significant and well-performed metric is MOtion-based Video Integrity Evaluation (MOVIE) index [13]. The MOVIE utilizes the optical flow estimation to adaptively guide spatial-temporal filtering with the three-dimensional (3-D) Gabor filter banks. The MOVIE index is proved to work the best in the LIVE Video Quality Database [14].

In this paper, we propose a systematic way to improve the existing video quality assessment method. Since MOVIE is known to perform very well as compared with other video quality metrics, we use it as the basic framework. Our idea can be simply sketched below. First, a video sequence is decomposed into smaller units, called the temporal decomposition unit (TDU). Then, the MOVIE index is used to compute quality scores in each TDU, which consists of a small number of frames. We choose the worst score in each TDU and adopt the regression approach to fuse the worst scores from all TDUs into a final score, which represents the ultimate quality of the input video. This proposed VQA model is called the temporal-decomposed MOVIE (TD_MOVIE). Another content-aware TD_MOVIE (CA-TD_MOVIE) metric with a variable TDU size selection mechanism based on the statistical property is also proposed to improve the performance furthermore.

The rest of this paper is organized as follows. Section II describes the proposed temporal-decomposed video quality metric (TD_MOVIE) with the worst score selection strategy. Next, we present performance results and comparisons with several relevant existing metrics in Section III. Finally, concluding remarks are given in Section IV.

II. PROPOSED VIDEO QUALITY ASSESSMENT METHOD

Since MOVIE is proved to work quite well on the evaluation of video contents [13], [14], we adopt it as the basic framework. The methodology presented in this paper is expected to be extended to other video quality metrics. Before introducing the proposed model, we will describe the MOVIE metric first.

A. MOVIE

MOVIE consists of two parts [13], which are Spatial MOVIE (S_{movie}) and Temporal MOVIE (T_{movie}), respectively. Consider a video sequence with *N* frames. Then, these two components of MOVIE are defined as follows:

$$S_{movie} = \frac{1}{N}\sum_{j=1}^{N} FE_s(t_j) , \quad (1)$$

$$T_{movie} = \sqrt{\frac{1}{N}\sum_{j=1}^{N} FE_T(t_j)} , \quad (2)$$

where $FE_S(t_j)$ and $FE_T(t_j)$ denote the frame level error indices for both spatial and temporal components of MOVIE at frame t_j. The overall MOVIE index (ST_{movie}) is then the product of (1) and (2):

$$ST_{movie} = S_{movie} \times T_{movie} . \quad (3)$$

B. Temporal Decomposition

First, a video is decomposed into smaller units along the temporal domain, called the temporal decomposition unit (TDU). Next, MOVIE is used to compute the quality score for each frame in the TDU. To be able to obtain the score of each frame, we need to modify (3) since it only can compute the score for the entire video sequence. The modified MOVIE frame score at frame t_j becomes

$$ST_{movie}(t_j) = FE_s(t_j) \times \sqrt{FE_T(t_j)} . \quad (4)$$

Suppose the video sequence is divided into M equal-length TDUs. Then we will have $n = \left\lceil \frac{N}{M} \right\rceil$ frames in each TDU, where $\lceil\cdot\rceil$ denotes the ceiling function. Let TDU_i denotes the *i*th TDU. The set of quality scores in TDU_i is

$$ST_{movie}(TDU_i) = \{ST_{movie}(t_{ij}) | \, j = 1, \cdots, n\} , \quad (5)$$

where $ST_{movie}(t_{ij})$ is the MOVIE quality score at the *j*th frame of the *i*th TDU.

C. The Worst Score Selection Strategy

According to the research conducted by sociologists, it is more likely for people to remember the unpleasant experience than the pleasant one [15]. This pattern can also be applied to what happened to the human perceptual quality of images and videos. When a test subject viewed a video, the most distorted video segment would attract the biggest attention from viewers, and this unpleasant experience is not easy to forget. Thus, the most distorted frame (e.g., the frame with the poorest score) impacts the human perception more than other frames. In other words, people tend to ignore the slightly distorted frames whenever the highly distorted frame presents. Since the worst video frame has the maximum MOVIE score, the score of temporal-decomposed MOVIE (TD_MOVIE) in the *i*th TDU is selected as

$$ST_{TD\_MOVIE}(i) = max\{ST_{movie}(t_{ij}) | \, j = 1, \cdots, n\} , \quad (6)$$

which is called the worst score selection strategy.

D. Fusion of Scores

To fuse the scores from each TDU into one final score, we have tried quite a few linear and nonlinear regression methods and found that the simple linear regression method yields a reasonably good result. Thus, it is adopted in this work. Consider the fusion of scores

$$\{ST_{TD\_MOVIE,k}(i) | \, i = 1, \cdots, M\}$$

for video *k*. Then, the TD_MOVIE quality score for the *k*th video is defined as

$$TD\_MOVIE_k = a_0 + \sum_{i=1}^{M} a_i * ST_{TD\_MOVIE,k}(i) . \quad (7)$$

In the training stage, we want to find the constant term a_0 and weighting coefficients a_i to minimize the difference between TD_MOVIE_k and the differential mean opinion score ($DMOS_k$). Namely,

$$\min_{a_0, a_i} |TD\_MOVIE_k - DMOS_k|, \forall k = 1, \cdots, n_1 , \quad (8)$$

where |. |denotes the l_1 norm and n_1 is the number of training videos. Here, $DMOS_k$ also represent the ground truth since they are obtained by human observers. Once a_0 and a_i's are decided, we can use (7) to compute the quality score TD_MOVIE_l for the *l*th testing video, where $l = 1, \cdots, n_2$ and where n_2 represents the number of testing videos.

To make sure the result stay unbiased, we use the *n*-fold cross-validation [16] to choose the training and testing video sets. For example, we divide the entre video database into *n* sets. Only one out of *n* sets is used for testing, and the remaining sets are used for training. Then, we perform this procedure *n* time, where each video set is used as the testing set once.

E. Content-Aware TD_MOVIE (CA-TD_MOVIE)

To achieve better correlation with human subjective scores, we use one index based on the statistics distribution of frame quality scores to classify all input videos into two groups before temporal decomposition. The statistical index d_i for the *i*th video is defined as

$$d_i = \frac{standard \;\; deviation \, \{ST_{movie}(t_j) | j = 1, \cdots, N\}}{mean \, \{ST_{movie}(t_j) | j = 1, \cdots, N\}} . \quad (9)$$

Intuitively, (9) represents the normalized variation of frame quality scores and it can be treated as one attribute of the underlying video content. Then, we select two different TDU sizes for these two groups of videos. This is called the content-aware TD_MOVIE and denoted by CA-TD_MOVIE, since the decomposition along the temporal domain (i.e., the number of TDUs) is based on a video content attribute, *i.e.,* d_i.

Theoretically, a larger d_i means a larger variation of frame score distribution. Consequently, we should choose a smaller TDU size, which yields a larger number of TDUs, in the decomposition. On the contrary, the group of videos with a smaller d_i value should be decomposed into a smaller number of TDUs. We will demonstrate the performance improvement with CA-TD_MOVIE in Section III.D.

III. Experimental Results

A. Test Database

To evaluate the performance of the proposed TD_MOVIE and CA-TD_MOVIE methods and other video quality metrics, we use the LIVE Video Quality Database [17]. It includes 10 reference videos; seven of them have a frame rate of 25 fps, while the other three have a frame rate of 50 fps. Besides, 15 test sequences are generated from each of the reference sequences using four different distortion processes. They are: simulated transmission of H.264 compressed bit streams through error-prone wireless networks and through error-prone IP networks, H.264 compression, and MPEG-2 compression. All videos have planar YUV 4:2:0 formats and 768×432 spatial resolutions. The subjective quality scores used in this database are differential mean opinion score (DMOS), ranging from 0 to 100.

B. Test Methodology and Performance Measure

Three indices are used to measure the performance of video quality metrics [6], [18]. The first index is the Pearson linear correlation coefficient (PLCC) between the objective and the subjective scores. It provides an evaluation of prediction accuracy. The second index is the Spearman rank order correlation coefficient (SROCC) between the objective and the subjective scores. It is considered as a measure of prediction monotonicity. The last index is the root-mean-squared error (RMSE), also between the objective and the subjective scores.

To account for the quality rating compression at the extremes of the test range, the following four-parameter, monotonic logistic function is used to fit the objective scores (VQA prediction) to the subjective scores (DMOS) before computing the first and second indices [6]:

$$f(x) = \frac{\beta_1 - \beta_2}{1 + e^{-(x-\beta_3)/|\beta_4|}} + \beta_2 , \tag{10}$$

where x is the objective score, $f(x)$ is the fitted objective score, and parameters β_j (j = 1,2,3,4) are chosen to minimize the least squares error between the subjective scores and the fitted objective scores. A better-performed quality metric should have higher PLCC, SROCC, and lower RMSE.

C. TD_MOVIE

We list the quality prediction performance of TD_MOVIE in Table I. In order to see the trend better, we also plot the PLCC performance in Fig. 1. As shown in Table I and Fig. 1, TD_MOVIE achieves the best performance when using five TDUs. Then, the performance goes down when using either a smaller or a larger number of TDUs.

TABLE I
PERFORMANCE MEASURE OF TD_MOVIE WHEN USING DIFFERENT NUMBER OF TDU

No. of TDUs Measure	1	2	3	4	5	6	7
PLCC	0.7791	0.8108	0.8110	0.8311	**0.8350**	0.8284	0.7934
SROCC	0.7755	0.8053	0.7994	0.8191	**0.8233**	0.8198	0.7883
RMSE	6.8811	6.4246	6.4215	6.1042	**6.0397**	6.1487	6.6823

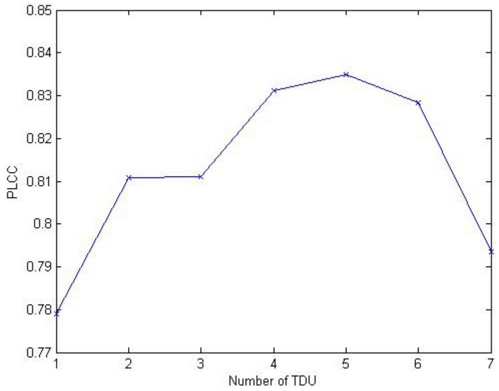

Fig. 1 PLCC performance of TD_MOVE

It is known that most of the videos in this database are 10 seconds long. Hence, each TDU corresponds to 2 seconds when a video is divided into five TDUs. This approximates the time frame for people to notice the quality change and make decision before forgetting what has been observed. In contrast, if one TDU is used, this TDU will correspond to 10 seconds. Since it lasts too long, it is probable for people to forget what they saw earlier. This explains why the proposed TD_MOVIE metric achieves better performance when using TDUs with duration of about 2 seconds. In other words, when the test video is longer than 10 seconds (e.g., 20 seconds), we need to divide it into more than five TDUs (e.g., 10 TDUs). On the contrary, less than five TDUs (e.g., 3 TDUs) will be used to decompose the video with a duration shorter than 10 seconds (e.g., 6 seconds).

TABLE II
PERFORMANCE MEASURE OF TD_MOVIE WHEN SELECTING DIFFERENT TYPE OF SCORES
(5 TDUs are used for fusion)

Type of Score Measure	Min.	Mean	Max.
PLCC	0.6990	0.8015	**0.8350**
SROCC	0.6581	0.7814	**0.8233**
RMSE	7.8504	6.5649	**6.0397**

To demonstrate that selecting the worst score would have the highest correlation with human perception as described in Section II.C, we also did the test on the fusion of the best score (i.e., minimum value of scores in (5)) and the medium score (i.e., mean value of scores in (5)), along with the worst score (i.e., maximum value of scores in (5)) in the experiment and show the results in Table II. As we can see from Table II, the worst score did offer the highest correlation with human subjective scores. Actually, the use of the worst score selection strategy has a substantial performance improvement over the mean and the best scores.

D. CA-TD_MOVIE

As mentioned in Section II.E, the CA-TD_MOVIE index is a content aware metric, which decides the TDU size (or the number of TDUs) used in temporal decomposition based on a video content attribute, denoted by d_i , which represents the normalized variation of frame scores. In the experiment, we decompose a video clip into five and four TDUs for the group of videos that has a larger and a smaller d_i values, respectively. Then, we repeat the TD_MOVIE quality measure process and show the results in Table III.

By comparing Table II and Table III, we see a clear improvement of CA-TD_MOVIE over TD_MOVIE. The PLCC scores increase from 0.8350 to 0.8494, the SROCC scores increase from 0.8233 to 0.8420 and the RMSE scores reduce from 6.0397 to 5.7932.

TABLE III
PERFORMANCE MEASURE OF CA-TD_MOVIE

Measure	CA-TD_MOVIE
PLCC	**0.8494**
SROCC	**0.8420**
RMSE	**5.7932**

E. Performance Comparison

TABLE IV
COMPARISON OF THE PERFORMANCE OF VQA MODELS

Measure / VQA Model	PLCC	SROCC	RMSE
PSNR	0.5465	0.5205	9.1929
V-SSIM	0.6058	0.5924	8.7337
VQM	0.7695	0.7529	7.0111
Q_{SVR}	0.7924	0.7820	6.6908
MOVIE	0.8116	0.7890	6.4130
TD_MOVIE	0.8350	0.8233	6.0397
CA-TD_MOVIE	**0.8494**	**0.8420**	**5.7932**

Table IV lists the performance of all VQA models, including state-of-the-art quality metrics and our two new metrics. Among all metrics in the comparison, Q_{SVR} [19] also belongs to the learning-oriented metrics [20]. The best performed metric is highlighted in bold in Table IV. It is clear that the two newly proposed metrics, TD_MOVIE and CA-TD_MOVIE, indeed improve the performance over MOVIE. Besides, they outperform other quality metrics in the table by a significant margin.

IV. CONCLUSION AND FUTURE WORK

In this work, we proposed a methodology to enhance the correlation performance of MOVIE by using temporal decomposition and selecting the worst scores for fusion. The worst score selection strategy was verified. Moreover, the results can be improved furthermore via adaptive TDU size selection based on a content aware mechanism. The methodology leads to new video quality metrics, called the TD_MOVIE and CA-TD_MOVIE. It was shown by experimental results that they both outperform MOVIE as well as other state-of-the-art video quality metrics by a significant margin.

As the next step of our current research, we will apply this temporal decomposition methodology to other quality metrics to see if the performance improvement is consistent. Also, we will conduct more experiments on other video databases to verify the robustness of the proposed methodology. If this works, then our new methodology will benefit other VQA metrics, not just for the MOVIE index.

REFERENCES

[1] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266-279, Apr. 2009.

[3] T.-J. Liu, K.-H. Liu, and H.-H. Liu, "Temporal information assisted video quality metric for multimedia," in *Proc. IEEE ICME*, pp. 697–702, Jul. 2010.

[4] M. Narwaria, W. Lin, "Machine Learning Based Modeling of Spatial and Temporal Factors for Video Quality Assessment," *IEEE ICIP*, 2011.

[5] Z. Wang, L. Lu, A. C. Bovik, "Video quality assessment using structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[6] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I," Mar. 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI.

[7] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception." *J. Opt. Soc. Am. A - Opt. Image Sci. Vis.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[8] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception." *Nat. Neurosci.*, vol. 9, no. 4, pp. 578–585, Apr. 2006.

[9] VQEG FRTV Phase I Database, 2000. [Online]. Available: ftp://ftp.crc.ca/crc/vqeg/TestSequences/.

[10] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[11] "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," *Recommendation ITU-R BT.1683*, Jan. 2004.

[12] "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," *Recommendation ITU-T J.144*, Feb. 2004.

[13] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. on Image Processing*, vol. 19, no. 2, pp. 335-350, Feb 2010.

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

[15] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *Proc. Int. Conf. Image Process.*, pp. 385–388, 2008.

[16] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.

[17] LIVE Video Quality Database. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html.

[18] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," Aug. 2003. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII.

[19] M. Narwaria and W. Lin, "Video Quality Assessment Using Temporal Quality Variations and Machine Learning," *IEEE ICME*, 2011.

[20] T.-J. Liu, W. Lin, and C.-C. J Kuo, "Recent Developments and Future Trends in Visual Quality Assessment," *APSIPA ASC*, Oct. 2011.