TEMPORAL INCONSISTENCY MEASURE FOR VIDEO QUALITY ASSESSMENT

Songnan Li, Lin Ma, Fan Zhang, King Ngi Ngan

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR

ABSTRACT

Visual quality assessment plays a crucial role in many vision-related signal processing applications. In the literature, more efforts have been spent on spatial visual quality measure. Although a large number of video quality metrics have been proposed, the methods to use temporal information for quality assessment are less diversified. In this paper, we propose a novel method to measure the temporal impairments. The proposed method can be incorporated into any image quality metric to extend it into a video quality metric. Moreover, it is easy to apply the proposed method in video coding system to incorporate with MSE for rate-distortion optimization.

Index Terms— video quality assessment, spatial visual quality measure, temporal inconsistency measure.

1. INTRODUCTION

A reliable objective Video Quality Metric (VQM) will benefit the video coding community at least in two ways. First, it can substitute for the cumbersome, slow and expensive subjective testing for the performance evaluation of different video coding approaches. Second, different from the subjective testing which must work in an off-line manner, the objective VQM can be embedded into video codec to guide the rate-distortion optimization.

Simple mathematical error measures, such as Mean Square Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), are widely used in video coding schemes to assess the visual quality of the encoded frames. However, it has been well acknowledged that these pixel-based difference measures do not correlate well with the perception of the human observers, because they operate on a pixel-by-pixel basis without considering the characteristics of the Human Visual System (HVS). Therefore, during the decades great efforts have been made towards accurate visual quality metrics, and progresses have been reported. However, although these advanced metrics can outperform MSE/PSNR in matching the subjective ratings, their complexity makes them impractical for replacing MSE/PSNR in the video coding system.

In this paper, we propose a Temporal Inconsistency Measure (TIM) for video quality assessment. Different from a stand-alone VQM, the proposed TIM can be incorporated into any Image Quality Metric (IQM) for its extension into a VQM. As we show in the sequel, it is very convenient to apply TIM in the video coding system to enhance the MSE for video quality assessment. The rest of this paper is organized as follows. Section 2 briefly reviews the related works, especially discussing the existing methods of using temporal information for video quality assessment. In Section 3, we propose TIM, and illustrate the cooperation of MSE and TIM (MSE_TIM) in measuring video quality. Section 4 gives the experimental results showing the performance of the MSE_TIM in matching the subjective ratings. Section 5 contains the concluding remarks.

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused Investment Scheme (Project 1903003).

2. RELATED WORKS

Many image quality metrics (IQM) can be generalized as:

$$F = \frac{\sum_{i=1}^{M} Q_i}{M} \tag{1}$$

where Q_i is the predicted quality value given to the local image patch indexed by i. Actually, (1) only defines the common spatial pooling method, whose objective is to calculate a single quality prediction for the whole image, given the quality score Q_i for each local image patch. The essence of each IQM lies in its particular definition of Q_i. For example, MSE quantifies the quality/distortion of an 8×8 image block by:

$$MSE = \frac{\sum_{k=1}^{64} (x_k - y_k)^2}{64} \tag{2}$$

where x and y are the reference block and the distorted block, respectively; k is the pixel index. Limited by the paper length, please refer to [1] etc. for a comprehensive review on IQM.

Video sequences consist of 2-D frames. Hence the simplest way to extend IQM into VQM is by using the following temporal pooling equation:

$$S = \frac{\sum_{n=1}^{N} F_n}{N} \tag{3}$$

where n is the frame index; N is the total number of the frames; F_n is the quality prediction for the n^{th} frame; S is the quality prediction for the video sequence. Eq. (3) cannot cope well with special temporal artifacts, e.g., jerky or jitter motions etc., yet it has been acknowledged that (3) can properly

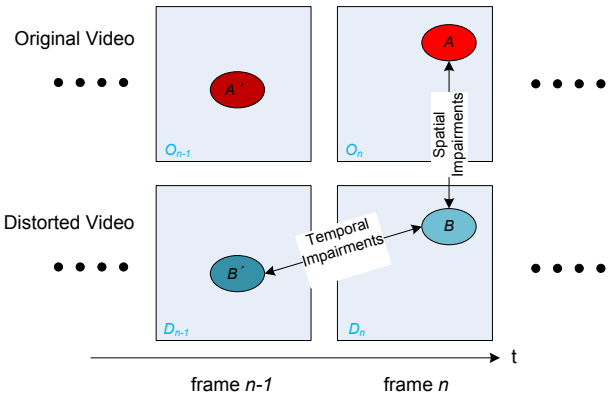


Fig. 1. Spatial and temporal impairments.

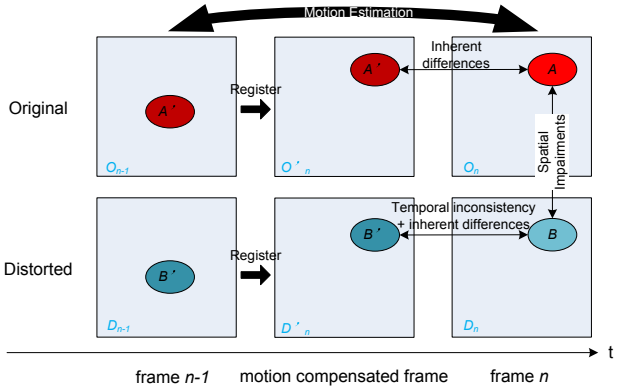


Fig. 2. Register frame $n-1$ to frame n.

handle the video coding artifacts, and it is a de facto method adopted to measure the visual quality of encoded video sequence, possibly ever since the advent of the video coding technique. Nevertheless, it is apparent that (3) does not exploit any temporal information which by proper use will benefit the visual quality assessment.

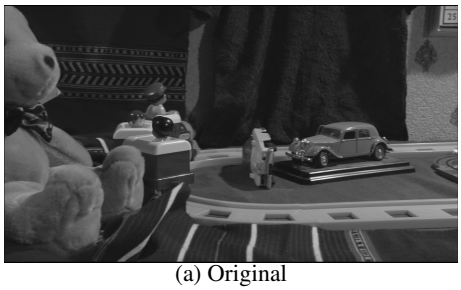
In the existing literature of video quality assessment, many metrics utilized temporal information to advance their prediction. For example, classical HVS-model based metrics [2] simulated the mechanism of the HVS to separate the visual signals into two temporal channels, one transient (bandpass) channel and one sustained (lowpass) channel. Both channels were further decomposed into multiple oriented spatial frequency sub-channels, each of which would be weighted according to the spatio-temporal contrast sensitivity function of the HVS. Metrics in [3] utilized rudimentary temporal information simply by differencing adjacent frames (without motion compensation). Bigger inter-frame difference leads to larger temporal masking effect. In Video Quality Model (VQM) [4], the absolute difference between adjacent frames was used to measure the moving-edge artifacts. Also in VQM, a varied of temporal pooling functions were employed, some of which took into consideration the temporal distribution of the impairments, e.g., by using the 90% most seriously distorted frames for the quality measure. In [5], smaller weightings were given to the large motion frames, since the authors found that their metric performs less stable when very large global motion occurs. The motion level of the frame was determined by the average motion vector length. The MOVIE index [6] utilized Gabor filters to decompose both reference and distorted video sequences into spatio-temporal 3-D frequency subbands. Spatial MOVIE index accumulated the differences within each subband to assess the spatial impairments; while temporal MOVIE index assigned different weightings to different subband differences for temporal impairment measure. The weighting value for each subband is inversely proportional to the distance between the subband and a spectral plane, the location of which is determined by the values of the spatial frequencies and the motion vectors. It is easy to see that the above-mentioned methods of utilizing temporal/motion information for visual quality measure are either too inconvenient to be incorporated into video coding or too simple to take full advantage of its usefulness.

3. TEMPORAL INCONSISTENCY MEASURE

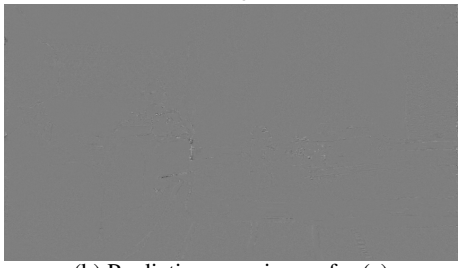
In Fig. 1, the eclipses represent the same object in different video frames. Due to the spatial impairments, A in the n^{th} frame of the original video (O_n) appears to be different from B in n^{th} frame of the distorted video (D_n); while due to the temporal impairments, B in D_n appears to be different from B' in D_{n-1}, i.e., the previous frame of D_n. In this paper, the term *temporal inconsistency* is used to describe the visual disparity of the same object in adjacent distorted frames. To the best of our knowledge, existing VQMs only used the temporal information, e.g., the differences of adjacent frames, motion vector length etc., to adjust the spatial impairments, but none of them attempt to measure the temporal inconsistency directly. We propose a Temporal Inconsistency Measure, abbreviated as TIM, in this work. The implementation issues of TIM and its cooperation with MSE are discussed in detail in this section.

3.1. Implementation

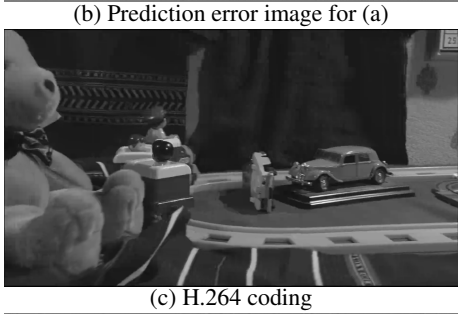
The crucial implementation issue of TIM is how to locate the same object in adjacent frames. To this end, we perform motion estimation on the original video sequence, and the obtained motion vectors are used to register frame O_{n-1} to O_n. The registration result is the so-called motion compensated frame (a term used in video coding) O'_n. Assuming that the distortions do not cause spatial/temporal shift or scaling of the object, which is true for many real-world distortions, the motion vectors derived from the original sequence, which are taken as the true representation of the motion trajectories of the objects, can be used for the distorted video to register D_{n-1} to D_n, generating the motion compensated frame D'_n. Next, by analyzing the differences between D_n and D'_n, which in fact can be treated as spatial impairments and mea-



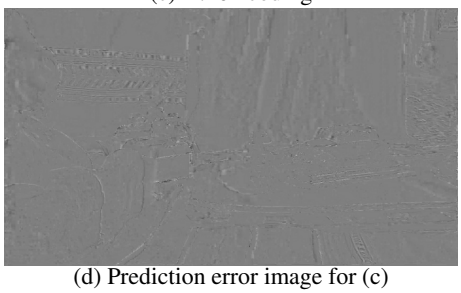
(a) Original



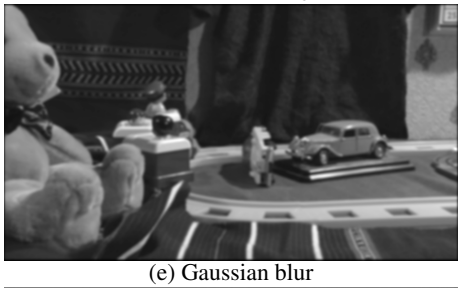
(b) Prediction error image for (a)



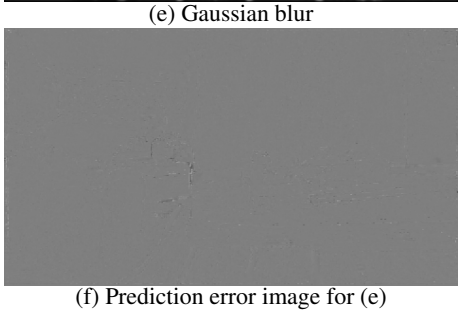
(c) H.264 coding



(d) Prediction error image for (c)



(e) Gaussian blur



(f) Prediction error image for (e)

Fig. 3. Prediction errors for different artifacts.

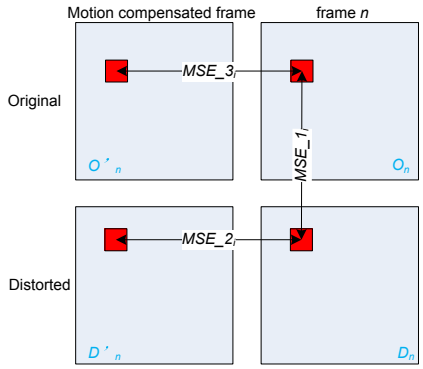


Fig. 4. TIM with MSE as the spatial impairment measure.

sured by any IQM, the temporal inconsistency can be quantified.

As shown in Fig. 3, (a), (c), and (e) are three n^{th} frames, from an original sequence, a H.264 encoded sequence, and a Gaussian blurred sequence, respectively. As mentioned above, we performed motion estimation (block-based full search) on the original sequence, and the derived motion vectors were used to predict (a), (c), and (e) from their previous frames. The prediction errors (differences of pixel values between the current frame and the motion compensated frame) are shown in Fig. 3 (b), (c) and (d). It can be observed that: (1) H.264 coded frame has larger prediction error, which implies larger temporal inconsistency; (2) Gaussian blurred frame has smaller prediction error. In video coding, the quality measure in use does not consider the temporal impairments at all. Therefore, video coding will cause large temporal artifacts especially when the compression ratio is high. On the other hand, in the Gaussian blurred sequence, each frame was filtered by the same Gaussian kernel, which intuitively will cause little temporal inconsistency. This is consistent with our observation in Fig. 3.

Perfect registration for the original sequence is impossible, because of, e.g., the restrictions imposed by the motion estimation algorithm, the changes of lighting condition, or the emergences of the occluded regions, and so on. Therefore, there will exist differences between A and A', as shown in Fig. 2 and Fig. 3 (b). We term it *inherent difference*. To measure the temporal inconsistency, we need to subtract the inherent differences from the total differences between B and B'. And because of the coexistence of the temporal inconsistency artifacts and the inherent differences, they mask each other visually. This masking effect should be considered by the quality metric.

3.2. MSE_TIM

Eq. (4) and Fig. 4 show how to incorporate MSE with TIM, where i is the index of an 8×8 block; MSE_1_i measures its spatial impairments; $max(MSE\_2_i - K \times MSE\_3_i, 0)$ measures its temporal impairments; K is a constant value which is larger than 1 to take account of the masking effect mentioned

$$MSE\_TIM = \frac{\sum_{i=1}^{M}(\omega \times MSE\_1_i + (1-\omega) \times max(MSE\_2_i - K \times MSE\_3_i, 0))}{M} \quad (4)$$

Table 1. Performance of the proposed video quality metric MSE_TIM.

		$\omega = 1$	$\omega = 0.8$	$\omega = 0.6$	$\omega = 0.4$	$\omega = 0.2$	$\omega = 0$
LIVE (coding)	LCC	0.457	0.467	0.481	0.504	0.578	0.651
	SROCC	0.420	0.433	0.448	0.459	0.523	0.625
	RMSE	9.142	9.092	9.014	8.877	8.384	7.804
LIVE (fullset)	LCC	0.570	0.603	0.614	0.632	0.640	0.675
	SROCC	0.552	0.559	0.574	0.591	0.630	0.667
	RMSE	9.016	8.756	8.665	8.505	8.427	8.097

above. The weighting value ω is to balance the importance of the spatial impairment measure and temporal inconsistency measure in the final quality decision. The experimental results of this metric MSE_TIM will be given in the next section.

To substitute MSE_TIM for MSE in video coding, we need to find the true motion vectors for each frame of the reference video. The rest of the computation payload will be low, since essentially the formulation of MSE (hereby its computational simplicity) is kept in MSE_TIM.

4. RESULTS

We tested MSE_TIM's performance on subjective video database LIVE [7], which includes 150 distorted videos generated from ten 768×432 reference videos with four different distortion types: H.264 coding, MPEG-2 coding, wireless transmission distortion, and IP transmission distortion. The so-called subjective video database provides each of its distorted video a subjective score to define its visual quality. These subjective scores are derived from subjective viewing tests where a large number of human observers participated and provided their opinions on the visual quality of each distorted video. Therefore, these subjective scores can be used as the ground truths to be compared with the objective scores given by a metric to evaluate its performance. After the non-linear mapping[1], three objective criteria were used to measure the correlation between the subjective scores and the nonlinearly mapped objective scores, which are the Linear Correlation Coefficient (LCC), the Spearman Rank-Order Correlation Coefficients (SROCC), and the Root Mean Squared Error (RMSE). Higher LCC and SROCC values indicate stronger correlation, i.e., better metric performance; while on the other hand, a smaller RMSE value indicates better metric performance.

Table 1 shows the performance of MSE_TIM which is formulized by (4). We used 8×8-block-based integer-pixel full search with the search range [-16, +16] as the motion estimation method; the constant K is experimentally set as 3. In Table 3, $\omega = 1$ correspond to the original MSE. As ω decreases, the influence of TIM on the final quality score increases. From the experimental results, it can be concluded that TIM helps in boosting MSE's performance on the coding artifacts and also the fullset of the database. Because of the tight correlation between the intensity of the spatial impairments and the temporal impairments, even better performance can be achieved when TIM is used only ($\omega = 0$).

[1]Limited by the paper length, for the motivation and implementation of the non-linear mapping please refer to [6] for reference.

5. CONCLUSION AND FUTURE WORK

We propose a temporal inconsistency measure which can be used with any IQM to assess video quality. For illustration, TIM is incorporated into MSE, and the resultant metric MSE_TIM is tested on the subjective video database LIVE. The experimental results show that TIM can improve MSE's performance in matching subjective ratings. Our future work will focus on investigating the influence of different motion estimation algorithms on TIM, employing MSE_TIM to guide video coding, and incorporating TIM into the state-of-the-art IQMs.

6. REFERENCES

[1] T.N. Pappas and R.J. Safranek, "Perceptual criteria for image quality evaluation", *Handbook of Image and Video Processing*, Academic Press: Orlando, FL, 2000.

[2] S. Winkler, "A perceptual distortion metric for digital color video", *Human Vision and Electronic Imaging IV*, 3644: pp. 175-184, 1999.

[3] W. Lin, "Gauging image and video quality in industrial applications", *Studies in Computational Intelligence* (SCI), 116: pp. 117-137, 2008.

[4] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", *IEEE Trans. on Broadcasting*, 50(3): pp. 312-322, 2004.

[5] Z. Wang, L. Lu, and A.C. Bovik, "Video quality assessment based on structural distortion measurement", *Signal Process.: Image Commun.*, 19(2): pp. 121-132, Feb. 2004.

[6] K. Seshadrinathan, A.C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos", *IEEE Trans. on Image Process.*, 19(2): pp. 335-350, Feb. 2010.

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