VIDEO QUALITY ASSESSMENT BY DECOUPLING ADDITIVE IMPAIRMENTS AND DETAIL LOSSES

Songnan Li, Lin Ma, King Ngi Ngan

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

ABSTRACT

In this paper, a review on existing methods of extending image quality metric to video quality metric is given. It is found that three processing steps are usually involved which include the temporal channel decomposition, temporal masking and error pooling. They are utilized to extend our previously proposed image quality metric, which separately evaluates additive impairments and detail losses, to video quality metric. The resultant algorithm is tested on subjective video database LIVE and shows a good performance in matching subjective ratings.

Index Terms**—** video quality assessment, distortion decoupling, human visual system, visual masking

1. INTRODUCTION

Since the human visual system (HVS) is the ultimate receiver of the video service, subjective viewing test is considered to be the most reliable way to evaluate visual quality. However, subjective viewing test is expensive, and not feasible for online manipulations, which makes it impractical for system design, quality monitoring, etc. Therefore, an accurate objective VQA algorithm, or namely video quality metric (VQM), becomes of fundamental importance to future multimedia applications.

It is customary to classify VQM into three categories according to the reference availability: full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics. In FR metrics, the reference is fully available and is assumed to have maximum quality. They can be applied in applications where the reference is fully available, such as image/video coding, watermarking etc. RR metrics extract features from the reference video, transmit them to the receiver side to compare against the corresponding features extracted from the distorted video. The design of RR metric mainly targets at quality monitoring. These features should be carefully selected to achieve both effectiveness and efficiency, i.e., predicting quality with great accuracy and small overhead for feature representation. NR metrics require no reference, therefore are most broadly applicable. For many no-reference applications, such as video signal acquisition, enhancement etc., NR metric is their only choice for on-line quality assessment. Not surprisingly, NR metric design is tough, facing challenges of limited input information. Therefore, to make sure acceptable prediction performance, many NR metrics are designed to cope with specific artifacts, such as blocking, blurring, ringing, jitter/jerky motion, etc., scarifying versatility for prediction accuracy. For a comprehensive overview on NR metrics, please refer to [1].

In this paper, we propose a FR video quality metric. It is an extension of our previously proposed image quality metric [2], which separately evaluates detail losses and additive impairments for visual quality assessment. In Section 2, we briefly review our IQM, and then discuss how to extend it to VQM. Section 3 elaborates the implementation details. Section 4 shows the performance of the proposed VQM in matching subjective ratings. Section V provides the concluding remarks.

2. BACKGROUND

2.1. Spatial distortion measurement

Limited by the paper length, please refer to [3] for an overview on image quality assessment. In our VQM, we adopt our previous work [2] to measure the spatial distortions. Instead of treating the spatial distortions indistinguishably, they are decomposed into details losses and additive impairments. As the name implies, detail losses refer to the loss of useful information which affects the content visibility. Additive impairments, on the other hand, refer to the redundant visual information which does not belong to the original image but appears in the distorted image. Their appearance will distract viewer's attention from the useful picture contents, causing unpleasant viewing experience. To assist understanding, an illustration is given in Fig. 1. In Fig. 1 (a), the distorted image is separated into the original image and the error image. Typically, HVS-model based IQMs will try to simulate low-level HVS responses to the error image, treating these distortions as being homogeneous. As shown in Fig. 1 (b), the proposed method will further separate the distortions into detail losses and additive impairments. For JPEG compressed images, as the one shown in Fig. 1, the

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

978-1-4577-1334-7/11/$26.00 ©2011 IEEE

additive impairment mainly appears to be blocky. In our implementation, we separate the distorted image into an additive impairment image and a restored image, as shown in Fig. 1 (c). The restored image exhibits the same amount of detail losses as the distorted image but is additive impairment free. Then, the detail loss can be obtained by subtracting the restored image from the original image. In [4], the necessity of decoupling linear frequency distortions and additive noises, two concepts essentially similar to detail losses and additive impairments, is firstly advocated and proved to be useful for visual quality assessment. In our viewpoint, the benefits of decoupling distortions into additive impairments and detail losses include several aspects. First, the content visibility can be more accurately quantified due to the extraction of detail losses. Second, a better spatial masking scheme can be designed. This will be explained in Section 3.4. Third, specific measurement can be developed to associate detail losses or additive impairments with visual quality. This will be explained in Section 3.5.

2.2. Extension from IQM to VQM

Reading the literatures, we found that three processing steps are usually involved for extension of IQM to VQM: temporal channel decomposition, temporal masking, and error/quality pooling.

Many existing VQMs decompose the video signal into multiple spatio-temporal frequency channels and then assign different weights to them according to, e.g., the contrast sensitivity function (CSF). It is believed that the early stage of the visual pathway separates visual information into two temporal channels: a low-pass channel and a band-pass channel, known as the sustained and transient channel, respectively. Several VQMs model this HVS mechanism by filtering the videos along the temporal dimension using one or two filters. Recently, Seshadrinathan et al. [5] proposed to use three-dimensional Gabor filters to decompose the video locally into 105 spatio-temporal channels enabling the calculation of motion vectors from the Gabor outputs. Different from typical CSF weighting, in [5] each channel is weighted according to the distance between its center frequency and a spectral plan identified by the motion vectors of the reference video. Lee et al. [6] proposed to find the optimal weights for channels by optimizing the metric's predictive performance on subjective video databases.

Masking is another visual phenomenon critical for video quality assessment: the visibility of distortions is highly dependent on both the local spatial and temporal activities. Lukas et al. [7] used the derivative of the outputs of a spatial visual model along the time axis to measure the local temporal activities, which then serves as input to a nonlinear temporal masking function. This function was calibrated by fitting psychophysical data. Lindh et al. [8] extended a classical divisive normalization based masking model from spatial to spatio-temporal frequency domain. Chou et al. [9] proposed to measure temporal activity simply by calculating pixel differences between adjacent frames. They constructed a temporal masking function via specifically designed psychophysical experiment. Similar temporal masking functions were taken by a host of video quality metrics and JND models.

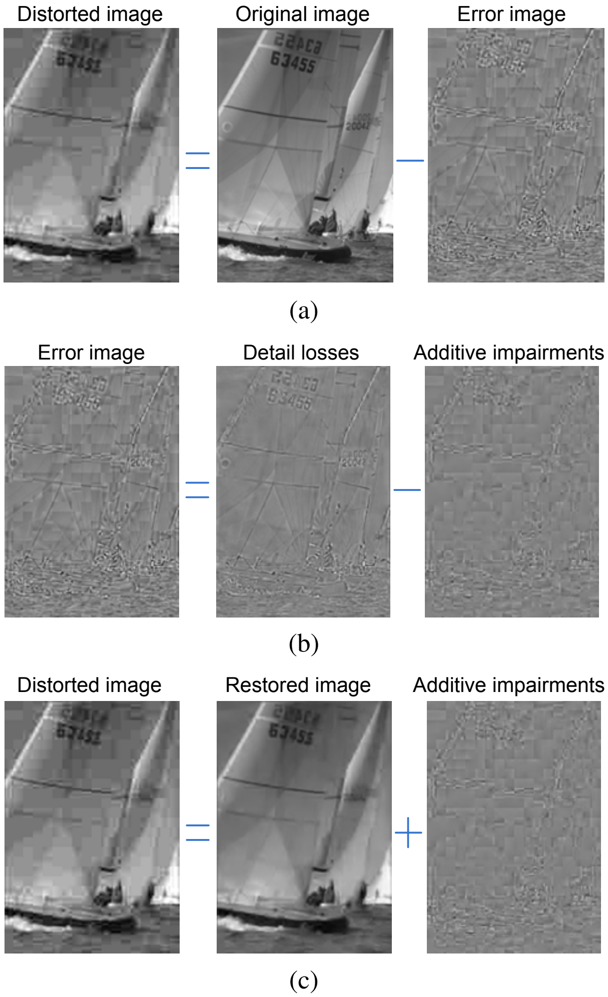

Fig. 1. An example of (a) separating the distorted image into the original image and the error image, (b) separating the error image into the detail loss image and the additive impairment image, (c) separating the distorted image into the restored image and the additive impairment image.

Pooling models the information integration which is believed to happen at the late stage of the visual pathway, and usually it is carried out by summation over all dimensions to obtain an overall quality score for an image or video. Wang et al. [10] used relative and background motions to quantify two terms: motion information content and perceptual uncertainty, which in the next step were used as weighting factors in the spatial pooling process. In TetraVQM [11], a degradation duration map is generated for each frame by analyzing the motion trajectory, and serves as a weighting matrix in spatial

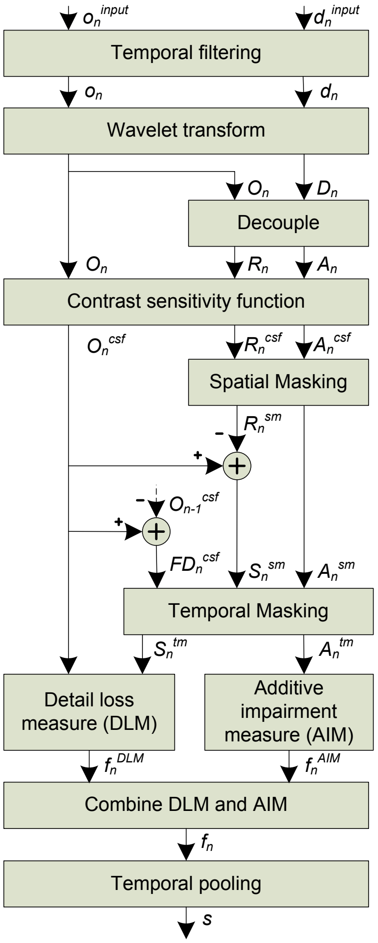

Fig. 2. The framework of the proposed VQM.

pooling. Ninassi et al. [12] proposed to take into account the temporal variation of spatial distortions in the temporal pooling process. They also considered the asymmetric human behavior in responding to quality degradation and improvement. This asymmetric human behavior was also modeled in [13].

We extend our IQM [2] to VQM by incorporating all the three processing steps mentioned above. To reduce computational complexity, the adopted decomposition, masking and pooling methods are simple and time efficient. In general, only the sustained channel is extracted after the temporal decomposition; temporal masking is calculated based on differences between adjacent frames; spatial pooling is formularized using Minkowski summation, and temporal pooling considers the aforementioned asymmetric human behavior. Implementation details will be given in the next section.

3. THE PROPOSED METHOD

The proposed video quality metric works with luminance only. Its framework is illustrated in Fig. 2. Detailed information on each processing component and the meaning of notations will be given below.

3.1. Temporal filtering

As introduced in Section 2, it is believed that there are two temporal channels in the HVS, a low-pass one, known as the sustained channel, and a band-pass one, known as the transient channel. However, it is the sustained channel that carries most visual information. In [14], it is confirmed that a majority of distortions exists in the sustained channel. Therefore as in [13] the proposed VQM uses the sustained channel only, to reduce the computational complexity. Specifically, both the original and distorted sequences are subjected to a 30Hz low-pass temporal filter. To reduce time delay, a three-tap infinite impulse response (IIR) filter is chosen [15]:

$$\mathbf{x}_n = 0.8 \times \mathbf{x}_n^{input} + 0.12 \times \mathbf{x}_{n-1}^{input} + 0.08 \times \mathbf{x}_{n-1} \qquad (1)$$

where $\mathbf{x}_n^{input}$ is either the n^{th} frame of the original sequence ($\mathbf{o}_n^{input}$) or that of the distorted sequence ($\mathbf{d}_n^{input}$), and $\mathbf{x}_n$ ($\mathbf{o}_n$ or $\mathbf{d}_n$) is the low-pass temporal filtering result.

3.2. Decoupling additive impairments and useful image contents

Each local patch $\mathbf{r}_i$ of the restored image can be decomposed into image components:

$$\mathbf{r}_i = \sum_{s=0}^{S} \mathbf{r}_i^s \qquad (2)$$

where i is the local index, $\mathbf{r}_i^s$ indicates the component reconstructed by the wavelet coefficients of the s^{th} subband (totally $S+1$ subbands), and $s = 0$ indicates the approximation subband. The same decomposition can be applied to the original image and the distorted image to derive $\mathbf{o}_i^s$ and $\mathbf{d}_i^s$ respectively. For $s \in \{1, ..., S\}$, the mean value of $\mathbf{r}_i^s/\mathbf{d}_i^s/\mathbf{o}_i^s$ equals zero. In general, we intend to get $\mathbf{r}_i^s$, $s \in \{1, ..., S\}$, that is additive-impairment-free and exhibits the same amount of detail losses as $\mathbf{d}_i^s$. To the first end, we make $\mathbf{r}_i^s = k_i^s \times \mathbf{o}_i^s$, where k_i^s is between 0 and 1 to take into account the influence of the detail loss. To the second end, we maximize the similarity between $\mathbf{r}_i^s$ and $\mathbf{d}_i^s$ by setting the scale factor k_i^s. The similarity between $\mathbf{r}_i^s$ and $\mathbf{d}_i^s$ is measured by the sum of squared differences to facilitate its optimization. Thus, the similarity maximization is implemented as: $min_{k_i^s \in [0,1]} ||\mathbf{r}_i^s - \mathbf{d}_i^s||^2$. Given an orthonormal discrete wavelet transform (DWT), the following equations hold:

$$\begin{array}{lll} & min_{k_i^s \in [0,1]} & ||\mathbf{r}_i^s - \mathbf{d}_i^s||^2 \\ = & min_{k_i^s \in [0,1]} & ||DWT[\mathbf{r}_i^s - \mathbf{d}_i^s]||^2 \\ = & min_{k_i^s \in [0,1]} & ||DWT[k_i^s \times \mathbf{o}_i^s - \mathbf{d}_i^s]||^2 \\ = & min_{k_i^s \in [0,1]} & ||k_i^s \times DWT[\mathbf{o}_i^s] - DWT[\mathbf{d}_i^s]||^2 \\ = & min_{k_i^s \in [0,1]} & ||k_i^s \times \mathbf{O}_i^s - \mathbf{D}_i^s||^2 \end{array} \qquad (3)$$

where $\mathbf{O}_i^s$ and $\mathbf{D}_i^s$ denote the DWT coefficients of $\mathbf{o}_i^s$ and $\mathbf{d}_i^s$ respectively. From (3), we can get the closed-form solution for the scale factor k_i^s:

$$k_i^s = clip(\frac{< \mathbf{O}_i^s \cdot \mathbf{D}_i^s >}{||\mathbf{O}_i^s||^2}, 0, 1) \qquad (4)$$

Simplification can be made that instead of using a vector of DWT coefficients, $\mathbf{O}_i^s$ or $\mathbf{D}_i^s$ can be represented by a single DWT coefficient. In this way, (4) is simplified to the division of two scalar values. In the following discussion, n index each frame, λ and θ index subband scale and orientation, respectively, and $\{i, j\}$ indexes the DWT coefficient position. A four-level $Haar$ DWT is applied to the temporally low-pass filtered original and distorted frames (o_n and d_n), generating the DWT coefficients $O_n(\lambda, \theta, i, j)$ and $D_n(\lambda, \theta, i, j)$. Based on the abovementioned simplification, scale factors of the high frequency subbands are given by:

$$k_n(\lambda, \theta, i, j) = clip(\frac{D_n(\lambda, \theta, i, j)}{O_n(\lambda, \theta, i, j) + 10^{-30}}, 0, 1) \qquad (5)$$

where the constant 10^{-30} is to avoid dividing by zero. Since intuitively the original mean luminance cannot be recovered from the distorted image, the approximation subband of the restored image is made to equalize that of the distorted image. Eventually, the DWT coefficients of the restored image can be obtained by:

$$R_n(\lambda, \theta, i, j) = \left\{ \begin{array}{ll} D_n(\lambda, \theta, i, j) & \theta = 1 \\ k_n(\lambda, \theta, i, j) \times O_n(\lambda, \theta, i, j) & otherwise \end{array} \right. \qquad (6)$$

where $\theta = 1$ indicates the approximation subband. Since DWT is a linear operator and the additive impairment image is given by $\mathbf{a}_n = \mathbf{d}_n - \mathbf{r}_n$, DWT coefficients of $\mathbf{a}_n$ can be calculated by:

$$A_n(\lambda, \theta, i, j) = D_n(\lambda, \theta, i, j) - R_n(\lambda, \theta, i, j) \qquad (7)$$

Notably, different from our previous work [2], the decoupling algorithm described in the paper cannot handle contrast enhancement, for the purpose of reducing computational complexity.

3.3. Contrast sensitivity function

HVS contrast sensitivity is the reciprocal of the contrast threshold, i.e., the minimum contrast value for an observer to detect a stimulus. It is found in psychovisual experiments that HVS contrast sensitivity depends on the characteristics of the visual stimulus, e.g., its spatial frequency, orientation, etc. Contrast sensitivity function (CSF) quantifies such dependences. The proposed VQM adopts the CSF used in [16]. It can be given by:

$$H(f, \theta) = \left\{ \begin{array}{ll} (a + bf_\theta)exp[-(cf_\theta)^{1.1}] & f \geq d \\ 0.981 & otherwise \end{array} \right. \qquad (8)$$

where f denotes the radial spatial frequency in cycles per degree of visual angle, and the constants are $a = 0.049$, $b = 0.592$, $c = 0.228$, $d = 3.4$. According to [17], the nominal spatial frequency of each DWT coefficient in scale λ can be given by:

$$f = \frac{\pi \times f_q \times d}{180 \times h \times 2^\lambda} \qquad (9)$$

1/30	1/30	1/30
1/30	1/15	1/30
1/30	1/30	1/30

Fig. 3. The weighting matrix **w**.

where d is the viewing distance, h is the picture height, and f_q is the cycles per picture height. In the following experiments, we set the ratio of d to h to be 6. $f_\theta = f/[0.15p(\theta) + 0.85]$ accounts for the oblique effect, i.e., the HVS is more sensitive to the horizontal and vertical channels than the diagonal channels. $p(\theta) = 1$ for the vertical and horizontal DWT subbands, and $p(\theta) = -1$ for the diagonal DWT subband. As illustrated in Fig. 2, we simulate CSF processing for the original image and the two decoupled images. It is implemented by multiplying each DWT coefficient with its corresponding CSF value derived from (8) and (9).

3.4. Spatial and temporal masking

Spatial masking refers to the visibility threshold elevation of a target signal caused by the presence of a superposed masker signal. Traditional spatial masking methods use original image to mask the distortions. However, artifacts may make the distorted image less textured compared to the original, especially for low-quality images where the contrasts of the textures or edges have been significantly reduced. In our metric, the restored image and the additive impairment image are decoupled, as illustrated in Fig.1 (c). Since the two decoupled images are superposed to form the distorted image, one's presence will affect the visibility of the other. Therefore, in the proposed metric both images serve as the masker to modulate the intensity of the other. We use (10) to calculate the spatial masking thresholds:

$$\mathbf{TH}_\lambda = m_i \times \sum_{\theta=1}^{3}(|\mathbf{M}_{\lambda,\theta}| \otimes \mathbf{w}), i \in \{s, t\} \qquad (10)$$

where **w** is a 3×3 weighting matrix as shown in Fig. 3, $|\mathbf{M}_{\lambda,\theta}|$ is the absolute DWT subband of the masker signal, operator $\otimes$ indicates convolution, and $\mathbf{TH}_\lambda$ is the spatial masking threshold map for each of the three DWT subbands in scale λ.The m_s can be used to alter the slope of the masking function. As in [2], m_s is set to 1 for all subbands. We take the absolute value of the CSF-weighted DWT coefficients, i.e., $|R_n^{csf}|$ ($|A_n^{csf}|$), subtract from them the spatial masking thresholds measured by (10) using A_n^{csf} (R_n^{csf}) as the masker, and clip the resultant negative values to 0. After the spatial masking, the DWT coefficients of the restored and additive impairment images can be represented by R_n^{sm} and A_n^{sm}, respectively. As shown in Fig.1, S_n^{sm} which denotes detail losses can be derived by subtracting R_n^{sm} from O_n^{csf}, i.e., the CSF-weighted DWT coefficients of the n^{th} original frame.

Temporal masking (TM) usually is modeled as a function of temporal discontinuity in intensity: the higher the inter-frame difference, the stronger the temporal masking effect. This method is also adopted by our VQM for computational simplicity. More precisely, the difference map between O_n^{csf} and O_{n-1}^{csf} is used as the masker to temporally mask the two types of spatial distortions: the detail losses (S_n^{sm}) and the additive impairments (A_n^{sm}). Eq. (10) is used to get the temporal masking threshold, and the masking process also follows the aforementioned three steps: taking absolute value, subtracting threshold and then clipping to zero. m_t is set to 0.5 for all subbands. The value is determined by training, which will be introduced in Section 4.

3.5. DLM, AIM and their combination

The additive impairment measure (AIM) and detail loss measure (DLM) are given by:

$$f_n^{AIM} = \frac{\sum_\lambda \sum_\theta [\sum_{i,j \in center} A_n^{tm}(\lambda, \theta, i, j)^2]^{1/2}}{N_p}, \theta \neq 1 \tag{11}$$

$$f_n^{DLM} = \frac{\sum_\lambda \sum_\theta [\sum_{i,j \in center} S_n^{tm}(\lambda, \theta, i, j)^2]^{1/2}}{\sum_\lambda \sum_\theta [\sum_{i,j \in center} O_n^{csf}(\lambda, \theta, i, j)^2]^{1/2}}, \theta \neq 1 \tag{12}$$

where $\theta \neq 1$ means that we exclude the use of approximation subband in spatial pooling, and $(i, j) \in center$ indicates that only the central region of each subband is used, which serves as a simple region of interest (ROI) model. Since additive impairments are relatively independent of the original image content, we assume that visual quality with respect to additive impairments can be predicted by analyzing their intensities without considering the original content. On the other hand, visual quality with respect to detail losses is supposed to be determined by the percentage of visual information losses. Therefore, in (11) and (12) the integrated distortion intensity is normalized by the pixel number N_p and the original image content, respectively. It should be noted that f_n^{DLM} is an approximate calculation of the percentage of visual information losses. Its low complexity makes the proposed metric time efficient.

f_n^{AIM} and f_n^{DLM} are combined by weighted summation:

$$f_n = w \times f_n^{AIM} + f_n^{DLM}; \tag{13}$$

The weighting factor w is determined by training, as will be introduced in Section 4. The training result is $w = 27.45$. Considering the typical range of f_n^{AIM} ($0 \sim 0.01$) and f_n^{DLM} ($0 \sim 0.3$), we can see that f_n^{AIM} and f_n^{DLM} are given similar weights.

3.6. Temporal pooling

We take the method used in [13] to perform the temporal pooling. In general, the asymmetric human behavior in responding to quality degradation and improvement is taken into account, that is, human observers are quick to criticize quality degradation and slow to response to quality improvement. It is achieved by using the following temporal pooling equations:

$$f_n^{'} = \begin{cases} f_{n-1}^{'} + a_- \triangle_n & if \quad \triangle_n \leq 0 \\ f_{n-1}^{'} + a_+ \triangle_n & if \quad \triangle_n > 0 \end{cases} \tag{14}$$

$$s = \frac{1}{N} \sum_{n=1}^{N} f_n^{'} \tag{15}$$

where $\triangle_n = f_n - f_{n-1}^{'}$. As in [13], we set a_-=0.04, and a_+=0.5.

4. EXPERIMENTS

In this section we present the predictive performance of the proposed VQM on subjective video database LIVE [18]. LIVE consists of 10 reference videos, and 150 test videos each of which is distorted by one of the four distortion types, i.e., H.264 compression, MPEG2 compression, wireless or IP transmission error, with various distortion intensities. The video resolution is 768×432. The database provides a subjective score (difference mean opinion scores, i.e., DMOS) for each of the distorted sequences. The subjective scores are derived from subjective viewing tests. They are taken as the ground truth to be compared with the metric outputs to evaluate the predictive performance. It is customary to nonlinearly map the metric scores to the ones that have a linear relationship with the subjective scores. And after the non-linear mapping[1], we use three objective criteria 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 a smaller RMSE value indicates better metric performance.

As mentioned in Section 3, there are two parameters, i.e., the temporal masking factor m_t and the weighting factor w, which are determined by training. The training set consists of 45 videos[2] from video database LIVE. The training objective is to maximize the SROCC value of the proposed VQM on the training set. The other 105 distorted videos are used for performance evaluation. The proposed VQM is compared with PSNR, a standardized metric Video Quality Model (VQ model) [19], and a state-of-the-art metric MOVIE [5]. As shown in Table 1 and 2, the proposed VQM demonstrates the best overall performance, and relatively good performance

[1] Limited by the paper length, please refer to [18] for the motivation and implementation of the non-linear mapping.

[2] The 45 distorted videos are generated from three reference sequences, i.e., *Station*, *Sunflower* and *Tractor*, randomly chosen from the LIVE video database.

Table 1. Overall performances on subjective video database LIVE (test set, 105 distorted sequences).

	PSNR	VQ Model	MOVIE	Proposed
LCC	0.485	0.796	0.828	**0.844**
SROCC	0.470	0.759	0.803	**0.835**
RMSE	9.863	6.820	6.216	**5.945**

Table 2. Performances (SROCC) on individual distortion types and distortion subsets (Coding:H.264+MPEG2, Trans.:IP+Wireless.

	PSNR	VQ Model	MOVIE	Proposed
H.264	0.490	0.760	0.856	**0.903**
MPEG2	0.357	**0.901**	0.830	0.866
IP	0.130	**0.774**	0.744	0.693
Wireless	0.584	0.760	0.801	**0.810**
Coding	0.392	0.777	0.833	**0.858**
Trans.	0.428	0.755	0.764	**0.822**

on the four individual distortion types and the two distortion subsets (Coding and Transmission Error). Regarding computational complexity, for 250-frame 768×432 sequences, PSNR, VQ Model and the proposed metric (all MATLAB implemented) require approximately 3 second, 1 minute, and 1 minute, respectively. MOVIE (C++ language implemented) needs much longer processing time, i.e., roughly 95 minutes.

5. CONCLUSION AND FUTURE WORK

In this paper, we extend our previous work [2], i.e., a image quality metric which separately evaluates detail losses and additive impairments for visual quality assessment, into a video quality metric, by incorporating three additional processing steps: temporal channel decomposition, temporal masking, and temporal pooling. Testing on subjective video database LIVE, the resultant VQM achieves a quite good performance in matching subjective ratings. In the future work, we will incorporate motion information into the abovemetioned three processing steps to more accurately model the HVS characteristics, and test the proposed VQM on more databases.

6. REFERENCES

[1] S.S. Hemami, and A.R. Reibman, ”No-reference image and video quality estimation: Applications and human-motivated design,” Signal Process.-Image Comm., vol. 25, no. 7, pp. 469-481, 2010.

[2] S.N. Li, L. Ma, K.N. Ngan, ”Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments,” IEEE Trans. on Multimedia, to appear, 2011.

[3] T.N. Pappas, and R.J. Safranek, ”Perceptual criteria for image quality evaluation,” Handbook of Image and Video Process., A. Bovik ed., Academic Press, 2000.

[4] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, ”Image quality assessment based on a degradation model,” IEEE Trans. on Image Process., vol. 9, no. 4, pp. 636-650, 2000.

[5] K. Seshadrinathan, and A.C. Bovik, ”Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos,” IEEE Trans. on Image Process., vol. 19, no. 2, pp. 335-350, 2010.

[6] C. Lee, and O. Kwon, ”Objective measurements of video quality using the wavelet transform,” Optical Engineering, vol. 42, no. 1, pp. 265-272, 2003.

[7] F.X.J. Lukas, ”Picture Quality Prediction Based on a Visual Model,” IEEE Trans. on Comm., vol. 30, no. 7, pp. 1679-1692, 1982.

[8] P. Lindh, and C.J.V.B. Lambrecht, ”Efficient spatio-temporal decomposition for perceptual processing of video sequences,” Proc. ICIP, vol. 3, pp. 331-334, 1996.

[9] C.H. Chou, and C.W. Chen, ”A perceptually optimized 3-D subband codec for video communication over wireless channels,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 2, pp. 143-156, 1996.

[10] Z. Wang, and Q. Li, ”Video quality assessment using a statistical model of human visual speed perception,” JOSA A-Optics, Image Science and Vision, vol. 24, no. 12, pp. B61-B69, 2007.

[11] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, ”Temporal Trajectory Aware Video Quality Measure,” IEEE Journal of Selected Topics in Signal Process., vol. 3, no. 2, pp. 266-279, 2009.

[12] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, ”Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment,” IEEE Journal of Selected Topics in Signal Process., vol. 3, no. 2, pp. 253-265, 2009.

[13] M. Masry, S.S. Hemami, and Y. Sermadevi, ”A scalable wavelet-based video distortion metric and applications,” IEEE Trans. Circuits and Syst. Video Technol., vol. 16, no. 2, pp. 260-273, 2006.

[14] S. Winkler. ”Quality metric design: A closer look,” In Proc. SPIE Human Vision and Electronic Imaging Conference, vol. 3959, pp. 3744, 2000.

[15] Z.H. Yu, H.R. Wu, ”Vision Model Based Digital Video Impairment Metrics,” Digital Video, Image Quality and Perceptual coding, H.R. Wu, and K.R. Rao eds., CRC Press, 2005.

[16] E.C. Larson, and D.M. Chandler, ”Most apparent distortion: full-reference image quality assessment and the role of strategy,” Journal of Electronic Imaging, vol. 19, no. 1, 2010.

[17] A.B. Watson, G.Y. Yang, J.A. Solomon, and J. Villasenor, ”Visibility of wavelet quantization noise,” IEEE Trans. Image Process., vol. 6, no. 8, pp. 1164-1175, 1997.

[18] K. Seshadrinathan and A. C. Bovik, ”Motion Tuned Spatio-temporal Quality Assessment of Natural Videos,” IEEE Trans. Image Process., vol. 19, no. 2, pp. 335-350, Feb. 2010.

[19] S. Wolf, and M.H. Pinson, ”Spatial-temporal distortion metrics for inservice quality monitoring of any digital video system,” in Proc. SPIE Conference on Multimedia Systems and Applications II, pp. 266-277, 1999.