Full-Reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments

Songnan Li, *Student Member, IEEE,* Lin Ma, *Student Member, IEEE,* and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—Video quality assessment plays a fundamental role in video processing and communication applications. In this paper, we study the use of motion information and temporal human visual system (HVS) characteristics for objective video quality assessment. In our previous work, two types of spatial distortions, i.e., detail losses and additive impairments, are decoupled and evaluated separately for spatial quality assessment. The detail losses refer to the loss of useful visual information that will affect the content visibility, and the additive impairments represent the redundant visual information in the test image, such as the blocking or ringing artifacts caused by data compression and so on. In this paper, a novel full-reference video quality metric is developed, which conceptually comprises the following processing steps: 1) decoupling detail losses and additive impairments within each frame for spatial distortion measure; 2) analyzing the video motion and using the HVS characteristics to simulate the human perception of the spatial distortions; and 3) taking into account cognitive human behaviors to integrate frame-level quality scores into sequence-level quality score. Distinguished from most studies in the literature, the proposed method comprehensively investigates the use of motion information in the simulation of HVS processing, e.g., to model the eye movement, to predict the spatio-temporal HVS contrast sensitivity, to implement the temporal masking effect, and so on. Furthermore, we also prove the effectiveness of decoupling detail losses and additive impairments for video quality assessment. The proposed method is tested on two subjective quality video databases, LIVE and IVP, and demonstrates the state-of-the-art performance in matching subjective ratings.**

***Index Terms*—Human visual system (HVS), spatial distortions decoupling, video quality assessment.**

I. Introduction

DEMANDS for various kinds of video services, such as digital TV, IPTV, video on demand, video conference, video surveillance, and so on, boom with the advances in video processing and communication technologies. A typical video service chain consists of sequential components, e.g., acquisition, enhancement, compression, transmission, transcoding, reconstruction, restoration, presentation, and so on, implemented by software and hardware systems. Video quality assessment (VQA) plays a central role in the life cycle of most of these systems: from the system design, performance evaluation, to the in-service functionality monitoring. In system design, an objective VQA algorithm can work online to bring benefits including performance improvement and/or resource saving; in performance evaluation, customer-oriented VQA can be employed to compare performances of competing systems, e.g., two different video coding schemes; for in-service quality monitoring, VQA continuously evaluates the system outputs, troubleshoots when system failure occurs, so as to guarantee high quality video service delivery. Since the human visual system (HVS) is the ultimate receiver of the video service, subjective VQA (subjective viewing test) is considered to be the most reliable way to evaluate visual quality. Its applications include assessing visual systems, benchmarking objective VQA algorithms, and so on. However, subjective VQA is not feasible for online manipulations, which makes it impractical for system design and quality monitoring. Furthermore, in order to ensure repeatable and statistically meaningful results, subjective VQA should precisely follow the standards [2]–[4] to set up the viewing environment, and should employ sufficient subjects in the viewing tests to account for individual differences. Satisfying these requirements makes subjective VQA time-consuming and expensive. Therefore, an accurate objective VQA algorithm, or namely video quality metric (VQM), becomes of paramount importance to the development of future video processing and communication applications.

The obstacle to an accurate VQM comes from numerous aspects: 1) the input video signal contents are diverse (e.g., sports, animations, movies, news report) creating problems, such as unpredictable visual attention, and so on; 2) visual systems generate various spatial artifacts (e.g., blocking, ringing, blurring, white noise) and/or temporal artifacts (e.g., jitter/jerky motion), which may be compounded during the video delivery; 3) the viewing conditions, such as the environment lightness, the display type and calibration, the viewing distance, and so on influence the distortion visibility; 4) the HVS is extremely complex and seems impossible to be completely modeled in the near future; and 5) visual quality judgment is viewer dependent, related to unpredictable factors such as the viewer's interests, expectation, quality experience, and so on. Consequently, VQM is not expected to be unconditionally as accurate and reliable as the subjective VQA. Instead, it is typical for most VQMs to restrict their utilities to specific sets of input contents, artifacts, and viewing conditions.

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

Manuscript received June 16, 2011; revised October 7, 2011; accepted December 11, 2011. Date of publication March 9, 2012; date of current version June 28, 2012. This paper was recommended by Associate Editor E. Magli.

The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: snli@ee.cuhk.edu.hk; lma@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2012.2190473

1051-8215/$31.00 © 2012 IEEE

FR metrics, the reference is fully available and is assumed to have maximum quality. They can be applied in system design and performance evaluation. For instance, in lossy video coding, the coding tools (e.g., intra and interprediction, transformation, quantization, de-blocking filter) and the coding modes (e.g., various intra or intermodes) are chosen according to their abilities to optimize rate and distortion, the latter of which can be quantified by FR metrics due to the availability of the reference. In watermarking, performances of various schemes are evaluated based on three major factors: robustness (to malicious attack), capacity (to embed information), and distortion perceptibility, where the distortion perceptibility can be measured by FR metrics. RR metrics extract features from the reference video, transmit them to the receiver side (by an ancillary channel [5] or by watermarking [6]) 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 inherently no-reference applications, such as video signal acquisition, enhancement, and so on, NR metric is their only choice for online quality assessment. Not surprisingly, NR metric design is tough, facing challenges of limited input information. Therefore, to guarantee acceptable prediction performance, many NR metrics are designed to cope with specific artifacts, such as blocking, blurring, ringing, jitter/jerky motion, and so on, scarifying versatility for prediction accuracy. For a comprehensive overview on NR metrics, please refer to [7].

In this paper, we propose a novel FR VQM. It promotes and completes our preliminary work [40] by extensively exploring the motion information and HVS temporal characteristics for video quality assessment. For spatial distortion measurement, we adopt the method [1] previously proposed by the authors for image quality assessment, i.e., separately evaluating detail losses and additive impairments which are decoupled from the overall spatial distortions. In Section II-A, we briefly overview FR image quality metrics (IQM) and explain in more detail the adopted spatial distortion measurement. In the following steps, we simulate the HVS processing, more specifically, how the human beings perceive the spatial distortions, using several major HVS characteristics, such as contrast sensitivity function (CSF), visual masking, information pooling, and so on. Compared with our preliminary work [40] in which motion information of the video was not considered and with most previous studies, the contribution of the proposed VQM comes from the extensive use of motion information to simulate the HVS processing, that is, motion vectors are derived in the wavelet domain, and employed in the eye-movement model, the spatio-velocity CSF, the motion-based temporal masking, and so on. We also simulate cognitive human behavior, which originally was proposed to be used in continuous quality evaluations [37], [38], and have proved its effectiveness in sequence-level quality prediction. In Section II-B, an introduction is given on how the previous VQMs use motion information and HVS characteristics for video quality assessment. The major differences between them and the proposed VQM are discussed in more detail. Section III elaborates the proposed VQM. Section IV shows the experimental results using both standard-definition and high-definition video databases to illustrate the performance of the proposed VQM in matching subjective ratings. Section V provides the conclusion.

II. Background

A. Full-Reference Image Quality Assessment

Most FR IQMs are general-purpose, that is, being able to handle various artifacts. As pointed out in [8], this cross-artifact versatility is crucial for benchmarking image processing systems. Pixel-based metrics, such as MSE/PSNR and so on, correlate visual quality with pixel value differences. As proved in many studies, MSE/PSNR predicts quality of white-noise distorted images with excellent accuracy, but failed to cope with other distortion types and cross-artifacts quality measurement. HVS-model-based metrics, such as those in [9]–[12], employ HVS models to simulate the low-level HVS processing of the visual inputs. They are often criticized, e.g., in [13], because these HVS models may be unsuitable for supra-threshold contrast measure and show a lack of geometric meaning [14, Fig. 4]. High-level HVS processing mechanism still cannot be fully understood. Therefore, many recent IQMs simply make use of common knowledge or assumption about the high-level HVS characteristics to guide quality prediction. For example, it is well acknowledged that structural information is critical for cognitive understanding; hence, the authors of [14] made the assumption that the structure distortion is a good representative of visual quality variation. They proposed the structure similarity index (SSIM) which distinguishes structure distortions from luminance and contrast distortions. This assumption has been well recognized and applied in other IQMs [8], [15], [16]. Another well-known assumption made by the authors of the visual information fidelity criterion (VIF) [17] is that the HVS correlates visual quality with the mutual information between the reference and distorted images. The mutual information resembles the amount of useful information that can be extracted from the distorted image. Although VIF seems to be quite different from SSIM in terms of the fundamental assumption, down to the implementation, the two IQMs share similarities, as analyzed in [18].

In this paper, we adopt our previous work [1] to measure the spatial distortions. Instead of treating the spatial distortions integrally, they are decomposed into detail 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. Their appearance in the distorted image 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 errors in a similar way. As shown in Fig. 1(b), the proposed method will

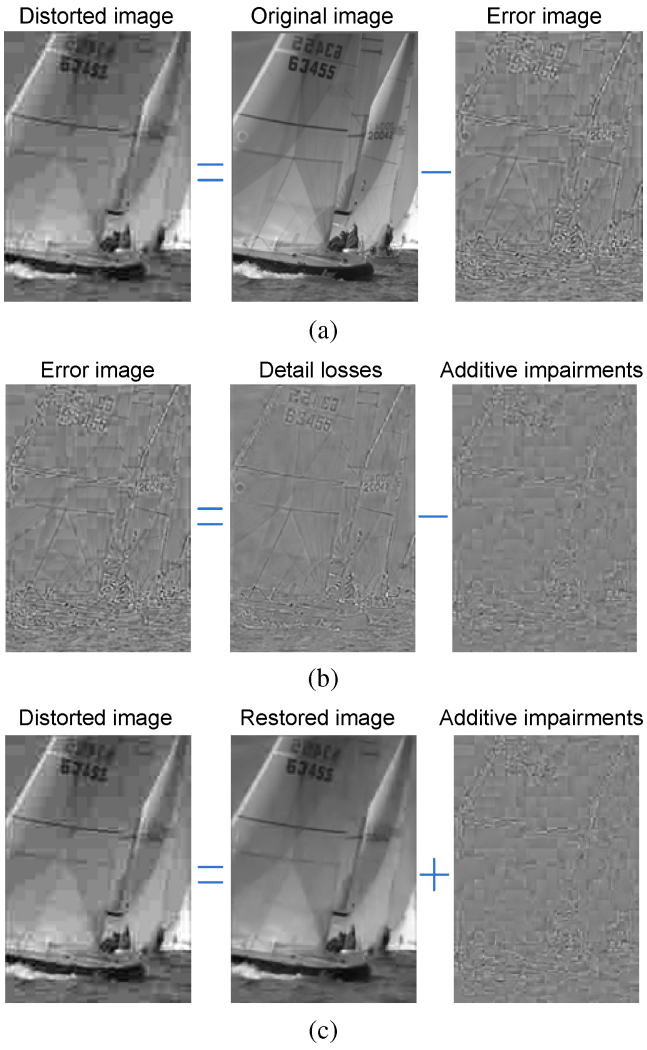


Fig. 1. Example of (a) separating the distorted image into the original image and the error image, (b) separating the error image into the detail loss image and the additive impairment image, and (c) separating the distorted image into the restored image and the additive impairment image.

further separate the image errors into detail losses and additive impairments. For JPEG compressed images, as the one shown in Fig. 1, the detail loss is caused by DCT coefficients quantization, and the additive impairment mainly appears to be blocky. In our implementation, we first 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. And then, the detail loss can be obtained by subtracting the restored image from the original one.

B. Exploiting Temporal Information for Video Quality Assessment

In comparison to images, videos involve one more dimension which apparently will render quality prediction more difficult. The predictive performance of the state-of-the-art VQMs measured by correlation coefficient is around 0.8 typically [19], while on the other hand for the state-of-the-art IQMs this value on most databases is above 0.9 [1]. To make video quality prediction more precise, besides accurately measuring the spatial distortions, motion information of the video and temporal characteristics of the HVS should be fully investigated.

A frequently used method of exploiting temporal HVS characteristics is to decompose the video signal into multiple spatio-temporal frequency channels and then assign different weights to them according to, e.g., the 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 channels, respectively. Several VQMs model this HVS mechanism by filtering the videos along the temporal dimension using one[1] or two filters. Recently, Seshadrinathan *et al.* [19] proposed to use 3-D Gabor filters to decompose the video locally into 105 spatio-temporal channels enabling the calculation of motion vectors from the Gabor outputs. Different from the typical CSF weighting, in [19] each channel is weighted according to the distance between its center frequency and a spectral plane identified by the motion vectors of the reference video. Lee *et al.* [20] proposed to find the optimal weights for channels by optimizing the metric's predictive performance on subjective quality video databases.

Masking is another visual phenomenon critical for video quality assessment. The visibility of distortions is highly dependent on both the spatial and temporal activities. However, most VQMs take into account only the spatial masking effect [19], [21]–[24]. And although temporal masking is involved in a few studies [25]–[27], motion vectors approximating the eye tracking movement are rarely investigated. Lukas *et al.* [25] 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.* [26] extended a classical divisive normalization-based masking model [28] from spatial to spatio-temporal frequency domain. Chou *et al.* [27] proposed to measure temporal activity simply by calculating pixel differences between adjacent frames. They constructed a temporal masking function via specifically designed psychophysical experiment. Chen *et al.* [29] conducted experiments to refine this function. Similar temporal masking functions were taken by a host of video quality metrics and JND models [30]–[33].

High-level temporal characteristics of the HVS can be used in the pooling process. 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.* [34] 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 [35], a degradation duration map is generated for each frame by analyzing the motion trajectory, and serves as a weighting matrix in spatial pooling. Ninassi *et al.* [36] proposed to take into account the temporal variation of spatial distortions in the temporal pooling process. In [24], [37], and [38], the authors

[1]For computational simplicity, only the sustained channel is isolated.

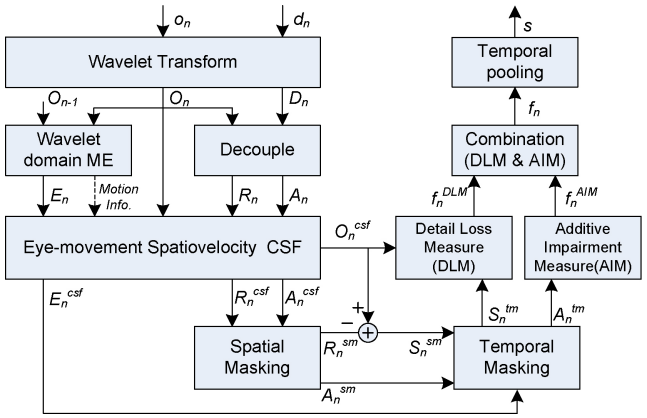


Fig. 2. Systematic framework of the proposed VQM.

considered the asymmetric human behavior in responding to quality degradation and improvement.

In this paper, three HVS processing simulation modules, i.e., the eye movement spatio-velocity CSF, the motion-based temporal masking, and the asymmetric temporal pooling, are deployed to extend our IQM [1] into the state-of-the-art VQM. Daly's eye movement spatio-velocity CSF [39] which models HVS sensitivity as a function of the spatial frequency and the velocity of the visual stimuli is used for the first time in video quality prediction. Compared to the other CSF models, Daly's method gives us mainly two advantages: 1) a more accurate CSF modeling; and 2) the involvement of the eye movement which is common in natural viewing conditions. We also propose a novel temporal masking method which differs from the most existing ones in the use of the motion information. The asymmetric temporal pooling method is directly adopted from [24] with the parameters tuned on a training set.

III. Proposed Method

A. Framework

The framework of the proposed VQM is illustrated in Fig. 2. It consists of several sequential processing modules. In general, for each frame of the distorted video, the distortions within are separated into additive impairments and detail losses. Motion-based contrast sensitivity function and visual masking are incorporated to rectify the intensities of the impairment images so that the intensity values become better representative of the low-level neural responses of the HVS. The influences of additive impairments and detail losses to visual quality are independently evaluated, and then combined together by weighted summation to generate quality score for each frame. An overall quality score indicating the quality of the whole video sequence is derived by temporal pooling over the individual quality scores of the video frames. Detailed information on each processing module and the meaning of the notations in Fig. 2 will be given below. It should be noted that the proposed VQM works with luminance only. Color inputs will be converted to gray scale before further processing.

B. Decoupling Additive Impairments and Useful Image Contents

As introduced in Section II-A, in our VQM each distorted video frame ($\mathbf{d}_n$, indexed by time n) is separated into an additive impairment image and a restored image, which is exemplified in Fig. 1(c). The decoupling algorithm which performs in the critically sampled wavelet domain is adopted from our previous work [1]. To reduce its computational complexity, two modifications are applied: 1) the original *db*2 wavelet transform is replaced by the simpler *Haar* wavelet transform; and 2) since it rarely happens in practical video applications, contrast enhancement is not distinguished from spatial distortions as in [1]. The derivation of the decoupling algorithm used in this paper is given below.

In the following context, $\mathbf{o}$, $\mathbf{d}$, $\mathbf{r}$, $\mathbf{a}$ are used to represent the original, distorted, restored, and additive impairment images, respectively. Our intention is to get the restored image $\mathbf{r}$ which exhibits the same amount of detail losses as the distorted image $\mathbf{d}$ but is additive impairment free. After calculating $\mathbf{r}$, the additive impairment image $\mathbf{a}$ can be obtained by $\mathbf{a} = \mathbf{d} - \mathbf{r}$. as shown in Fig. 1(c).

The restored image $\mathbf{r}$ consists of local image patches $\mathbf{r}_i$ where i indicates the local position. Each patch $\mathbf{r}_i$ can be further decomposed into multiple components of different spatial frequencies

$$\mathbf{r}_i = \sum_{s=0}^{S} \mathbf{r}_i^s \tag{1}$$

where s indicates the nominal spatial frequency. Such decomposition which requires both time and frequency localization of the signal can be implemented by wavelet transform. Therefore, $\mathbf{r}_i^s$ can be considered as the image component reconstructed from the wavelet coefficients of subband s around location i. The decomposition is applied to the original image $\mathbf{o}$ and the distorted image $\mathbf{d}$ to derive $\mathbf{o}_i^s$ and $\mathbf{d}_i^s$, respectively. To make $\mathbf{r}_i^s$ additive impairment free, the gradient of $\mathbf{r}_i^s$ should be proportional to that of $\mathbf{o}_i^s$, i.e., $\nabla \mathbf{r}_i^s = k_i^s \times \nabla \mathbf{o}_i^s$, where the value of k_i^s is limited to [0,1] to account for the detail losses. For high frequency subbands ($s \in \{1, 2, ..., S\}$), the mean values of $\mathbf{r}_i^s$ and $\mathbf{o}_i^s$ are equal to zero. Therefore, to satisfy the requirements on $\nabla \mathbf{r}_i^s$, we impose that $\mathbf{r}_i^s = k_i^s \times \mathbf{o}_i^s$.

Another objective of the decoupling is to make $\mathbf{r}_i^s$ exhibit the same amount of detail losses as $\mathbf{d}_i^s$. The distorted image patch $\mathbf{d}_i^s$ can be taken as the composite of two signals, i.e., the useful image content and the additive impairments. Since the additive impairments correlate poorly with $\mathbf{r}_i^s$, as exemplified by Fig. 1(b), the similarity between $\mathbf{r}_i^s$ and $\mathbf{d}_i^s$ is approximately equivalent to the similarity between $\mathbf{r}_i^s$ and the useful image content of $\mathbf{d}_i^s$. Therefore, making $\mathbf{r}_i^s$ exhibit the same amount of detail losses as $\mathbf{d}_i^s$ (i.e., maximizing the similarity between $\mathbf{r}_i^s$ and the useful image content of $\mathbf{d}_i^s$) can be achieved equivalently by maximizing the similarity between $\mathbf{r}_i^s$ and $\mathbf{d}_i^s$. For computational simplicity, the image similarity is measured by the sum of squared differences: $\min_{k_i^s \in [0,1]} ||\mathbf{r}_i^s - \mathbf{d}_i^s||^2$. Given an orthonormal discrete wavelet transform (DWT), the following equations hold:

$$\begin{aligned} & \min_{k_i^s \in [0,1]} && ||\mathbf{r}_i^s - \mathbf{d}_i^s||^2 \\ = \ & \min_{k_i^s \in [0,1]} && ||DWT[\mathbf{r}_i^s - \mathbf{d}_i^s]||^2 \\ = \ & \min_{k_i^s \in [0,1]} && ||DWT[k_i^s \times \mathbf{o}_i^s - \mathbf{d}_i^s]||^2 \\ = \ & \min_{k_i^s \in [0,1]} && ||k_i^s \times DWT[\mathbf{o}_i^s] - DWT[\mathbf{d}_i^s]||^2 \\ = \ & \min_{k_i^s \in [0,1]} && ||k_i^s \times \mathbf{O}_i^s - \mathbf{D}_i^s||^2 \end{aligned} \tag{2}$$

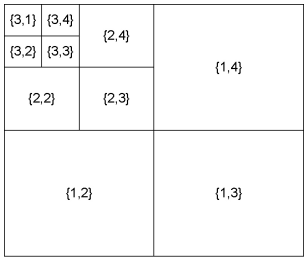


Fig. 3. Subband indexing. Each subband is indexed by a level and an orientation $\{\lambda, \theta\}$. $\theta = 2$, 3, and 4 denote the vertical, diagonal, and horizontal subbands, respectively.

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. The closed-form solution of the scale factor k_i^s in (2) can be given by

$$k_i^s = clip\left(\frac{<\mathbf{O}_i^s \cdot \mathbf{D}_i^s>}{||\mathbf{O}_i^s||^2}, 0, 1\right) \tag{3}$$

where $< \cdot >$ denotes the inner product operation, and $clip(x, 0, 1)$ is equivalent to $\min(\max(0, x), 1)$. According to the size of the local image patch, $\mathbf{O}_i^s$ and $\mathbf{D}_i^s$ can be a vector of DWT coefficients or just a single DWT coefficient. We found experimentally that the patch size is inessential to the decoupling performance. Therefore, for computational simplicity we use a single DWT coefficient to represent $\mathbf{O}_i^s$ and $\mathbf{D}_i^s$. In this way, (3) is simplified to be the division of two scalar values.

In the following discussion, n is used to index each frame, $\{\lambda, \theta\}$ is used to index each wavelet subband, as illustrated in Fig. 3, and $\{i, j\}$ is used to index the DWT coefficient position. A four-level *Haar* DWT is applied to the original ($\mathbf{o}_n$) and distorted ($\mathbf{d}_n$) frames, generating the DWT coefficients $O_n(\lambda, \theta, i, j)$ and $D_n(\lambda, \theta, i, j)$, respectively. Based on the above analysis, scale factors of the high frequency subbands can be given by

$$k_n(\lambda, \theta, i, j) = clip\left(\frac{D_n(\lambda, \theta, i, j)}{O_n(\lambda, \theta, i, j) + 10^{-30}}, 0, 1\right) \tag{4}$$

where the constant 10^{-30} is to avoid dividing by zero. Since intuitively the original mean luminance cannot be recovered from the distorted frame, the approximation subband of the restored image is made to equal that of the distorted one. Eventually, the DWT coefficients of the restored image $R_n(\lambda, \theta, i, j)$ can be obtained by

$$R_n(\lambda, \theta, i, j) = \begin{cases} D_n(\lambda, \theta, i, j), & \theta = 1 \\ k_n(\lambda, \theta, i, j) \times O_n(\lambda, \theta, i, j), & \text{otherwise} \end{cases} \tag{5}$$

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

$$A_n(\lambda, \theta, i, j) = D_n(\lambda, \theta, i, j) - R_n(\lambda, \theta, i, j). \tag{6}$$

C. Motion Estimation

Motion estimation (ME) algorithms aim at finding motion vectors representing object trajectory. In video compression, motion vectors can be used to remove the interframe redundancy. In video quality assessment, as will be elaborated below, the motion of the visual stimuli affects the HVS sensitivity. The use of motion vectors can enhance the modeling accuracy of how the HVS perceives distortions. As shown in Fig. 2, the ME is performed in the wavelet domain of the original video sequence. Wavelet transform decomposes a video frame into subbands of different resolutions. The motion vectors of these multiresolution subbands are highly correlated since they actually specify the same motion structure at different scales. Therefore, Zhang *et al.* [41] proposed a wavelet domain multiresolution ME scheme, where motion vectors at higher resolution are predicted by the motion vectors at the lower resolution and are refined at each step. The proposed method not only considerably reduces the searching time, but also provides a meaningful characterization of the intrinsic motion structure. We directly adopt Zhang's scheme to perform the wavelet domain ME. Specifically, the motion vectors $V_{\lambda,\theta}$ for the coarsest subbands ($\lambda = 4, \theta = 1, 2, 3, 4$) are estimated by block-based integer-pixel full search with a search range of $[-3, +3]$. These motion vectors are then scaled by $V_{\lambda,\theta} = 2^{(4-\lambda)} \times V_{4,\theta}$ ($\lambda = 1, 2, 3, \theta = 2, 3, 4$), and used as initial estimates for the finer subbands. These initial motion vectors are further refined by using full search with a relatively small search range $[-2, +2]$. The block size is adapted to the scale, i.e., $2^{(5-\lambda)} \times 2^{(5-\lambda)}$ ($\lambda = 1, 2, 3, 4$), so that no interpolation is needed as the motion vectors propagate through scales. The motion vectors will be used in the contrast sensitivity function. The motion estimation errors, E_n, are also saved for use in the temporal masking. It is worth noting that the critically sampled wavelet domain ME is not as accurate as the spatial domain ME due to the shift-variant property caused by the decimation process of the wavelet transform. As will be discussed below, we try to partially compensate the ME imprecision in the temporal masking process. More advanced wavelet domain ME algorithms can be used, but we found Zhang's scheme a good compromise between performance and simplicity.

D. Contrast Sensitivity Function

Human contrast sensitivity is the reciprocal of the contrast threshold, i.e., the minimum contrast value for an observer to detect a stimulus. It is related to the properties of the visual stimulus, notably its spatial and temporal frequencies. A spatio-temporal CSF is a measure of the contrast sensitivity against spatial and temporal frequencies of the visual stimulus. In the proposed method, we adopt Daly's eye movement spatio-velocity CSF model [39], as shown in Fig. 4, which was extended from Kelly's spatio-temporal CSF model [42] by taking into account the natural drift, smooth pursuit, and saccadic eye movements. In natural viewing conditions, evaluating visual quality must involve eye movements. Hence, incorporating eye movement model in the VQM should help to boost its predictive performance. Furthermore, rather than using flickering waves, such as in [23] and [43], the derivation of Daly's model is based on traveling waves which better represent the natural viewing conditions [39], [44].

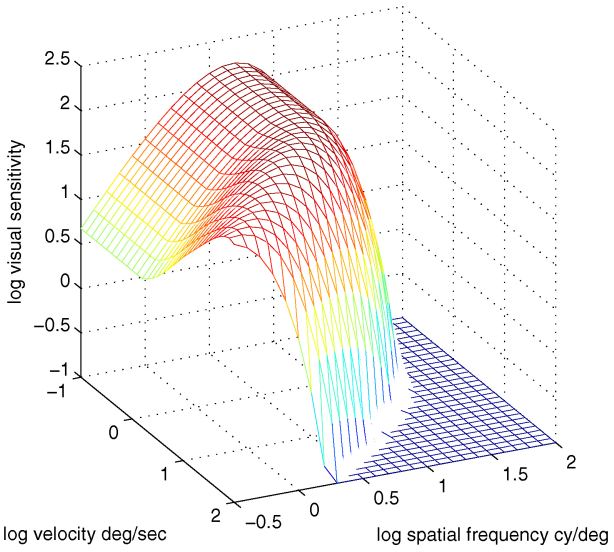


Fig. 4. Daly's eye movement spatio-velocity CSF model.

Daly's model can be formulated as

$$CSF(\rho, v_R) = k \cdot c_0 \cdot c_1 \cdot c_2 \cdot v_R \cdot (c_1 2\pi\rho)^2 exp(-\tfrac{c_1 4\pi\rho}{\rho_{\max}})$$
$$k = s_1 + s_2 \cdot |log(c_2 v_R/3))|^3 \quad (7)$$
$$\rho_{\max} = p_1/(c_2 v_R + 2)$$

where ρ is the spatial frequency of the visual stimulus in cy/deg and v_R is the retinal velocity in deg/s. The values of the constants are consistent with that of [39], i.e., $c_0 = 1.14$, $c_1 = 0.67$, $c_2 = 1.92$, $s_1 = 6.1$, $s_2 = 7.3$, $p_1 = 45.9$. Since they are not quite related to our subject, please refer to [39] for the physical meanings of these constants. The nominal spatial frequency of subbands in scale λ can be given by

$$\rho(\lambda) = \frac{\pi \cdot f_s \cdot d}{180 \cdot h \cdot 2^{\lambda}} \quad (8)$$

where d is the viewing distance, h is the picture height, and f_s is the cycles per picture height. The retinal velocity v_R can be calculated as follows:

$$v_R = v_I - v_E \quad (9)$$

where v_I is the image plane velocity, and v_E is the eye velocity. There are three types of eye movements modeled, i.e., the natural drift eye movements (0.8–0.15°/s) which are responsible for the perception of static imagery during fixation, the saccadic eye movements (160–300°/s) which are responsible for rapidly moving the fixation point from one location to another, and the smooth pursuit eye movements which lies between the two endpoints and occur when the eye is tracking a moving object. Adopted from [39], the equation used to model the eye velocity v_E as a function of the image plane velocity v_I is

$$v_E = \min\lfloor (g_{sp} \cdot v_I) + v_{\text{MIN}}, v_{\text{MAX}} \rfloor \quad (10)$$

where g_{sp} is the gain of the smooth pursuit eye movements modeling the eye tracking lag, v_{MIN} is the minimum eye velocity due to drift, and v_{MAX} is the maximum eye velocity before transmitting to saccadic movements. As in [39], we set $g_{sp} = 0.82$, $v_{\text{MIN}} = 0.15$, $v_{\text{MAX}} = 80$. Given the magnitude, v, of the motion vector and the frame rate, f_r, in frames per second (f/s), the image plane velocity can be calculated as

$$v_I = \frac{180 \cdot h \cdot 2^{\lambda} \cdot v \cdot f_r}{\pi \cdot d \cdot f_s} \quad (11)$$

where the meaning of the notations is the same as (8). Using (7)–(11) and the motion vectors derived from the wavelet domain ME, for each wavelet coefficient a CSF value can be calculated. As illustrated in Fig. 2, we simulate the CSF processing in the wavelet domain of the original image, the two decoupled images, and the ME prediction error image. It is implemented by multiplying each wavelet coefficient with its corresponding CSF value. The resultant signals are denoted as O_n^{csf}, R_n^{csf}, A_n^{csf}, E_n^{csf}, respectively.

E. 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 and edges have been significantly reduced. In our method, 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. As in [1], the equation used to calculate the spatial masking thresholds is

$$\mathbf{T}_{\lambda} = m_s \times \sum_{\theta=1}^{3} (|\mathbf{M}_{\lambda,\theta}| \otimes \mathbf{w}) \quad (12)$$

where $\mathbf{w}$ is a 3×3 weighting matrix with the central element being 1/15 and all the other elements being 1/30, $|\mathbf{M}_{\lambda,\theta}|$ is the absolute value of the $\{\lambda, \theta\}$ wavelet subband of the masker signal, $\otimes$ denotes convolution, and T_{λ} is the threshold map for the three wavelet subbands in scale λ. The m_s determines the slope of the masking function. As in [1], we set $m_s = 1$ for all scales. We take the absolute value of the CSF-weighted wavelet coefficients of the maskee signal, i.e., $|R_n^{csf}|$ ($|A_n^{csf}|$), subtract from them the spatial masking thresholds given by (12) with A_n^{csf} (R_n^{csf}) as the masker signal, and clip the resultant negative values to 0. After the spatial masking, the wavelet coefficients of the restored and additive impairment images are represented by R_n^{sm} and A_n^{sm}, respectively. As shown in Fig. 2, S_n^{sm} which denotes detail losses is derived by subtracting R_n^{sm} from O_n^{csf}, i.e., the CSF-weighted wavelet coefficients of the nth original frame.

As aforementioned in Section II, temporal masking is often modeled as a function of temporal discontinuity, that is, the higher the interframe differences, the stronger the temporal masking effect. To the best of our knowledge, temporal masking methods in the literature, such as the ones used in [27] and [30]–[33], measure interframe differences of the original video sequence ($\mathbf{o}_n - \mathbf{o}_{n-1}$) as the masker to mask distortions,

and notably they do not consider the eye movement. As discussed above, the human eyes may track the motion. Hence, the use of interframe differences without taking into account motion vectors may exaggerate the temporal masking effect. In our implementation, the CSF-weighted ME prediction error image, E_n^{csf}, is used to calculate the temporal masker. In case of inaccurate motion vectors as discussed in Section III-C, the temporal masker is given by $\min(E_n^{csf}, (O_n^{csf} - O_{n-1}^{csf}))$, which means that each element of the temporal masker equals the smaller one of the prediction error and the interframe difference. Equation (12) with the spatial masking factor m_s replaced by the temporal masking factor m_t is used to calculate the temporal masking thresholds. We set $m_t = 0.4$ by tuning the predictive performance of the VQM on a training set, as to be explained in Section IV. The masking process also follows the aforementioned three steps: taking absolute value of the maskee, subtracting masking thresholds, and then clipping negative values to zero. The temporal masker is used to modulate the intensities of the two types of spatial distortions: the detail losses (S_n^{sm}) and the additive impairments (A_n^{sm}). The resultant signals are denoted as S_n^{tm} and A_n^{tm}, respectively.

F. Two Quality Measures and Their Combination

As in [40], the additive impairment measure (AIM) and the detail loss measure (DLM) are given by

$$f_n^{\mathrm{AIM}} = \frac{\sum_\lambda \sum_\theta [\sum_{i,j\in \mathrm{center}} A_n^{tm}(\lambda,\theta,i,j)^2]^{1/2}}{N_p} \quad \theta \neq 1 \tag{13}$$

$$f_n^{\mathrm{DLM}} = \frac{\sum_\lambda \sum_\theta [\sum_{i,j\in \mathrm{center}} S_n^{tm}(\lambda,\theta,i,j)^2]^{1/2}}{\sum_\lambda \sum_\theta [\sum_{i,j\in \mathrm{center}} O_n^{csf}(\lambda,\theta,i,j)^2]^{1/2}} \quad \theta \neq 1 \tag{14}$$

where $\theta \neq 1$ means that we exclude the use of approximation subband in the 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 model. Since additive impairments are relatively independent of the original 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 connected with the percentage of visual information loss. Therefore, in AIM and DLM the integrated distortion intensities are normalized by the pixel number N_p and an integrated value associated with the original content,[2] respectively.

To derive the frame-level quality score, f_n^{AIM} and f_n^{DLM} are combined by weighted summation[3]

$$f_n = f_n^{AIM} + w \cdot f_n^{DLM}. \tag{15}$$

The weighting factor w is set to 2.47×10^3 by performance tuning, as to be explained in Section IV.

[2]To make complexity affordable, DLM only provides an approximate calculation of the visual information loss.

[3]The typical value ranges of f_n^{ALM} and f_n^{DLM} are from 0 to 600 and from 0 to 0.3, respectively.

G. Temporal Pooling

In temporal pooling, quality scores f_n, $n \in \{1, ..., N\}$, of the video frames are integrated to yield the overall quality score of the whole video sequence. Rather than directly calculating their average, preprocessing on the frame-level scores is performed to generate intermediate results f_n', $n \in \{1, ..., N\}$. The preprocessing simulates several cognitive human behaviors that have been reflected in the results of many continuous quality evaluations [37], [38], like the smoothing effect, i.e., the subjective ratings typically demonstrate far less variation than the objective quality scores, and the asymmetric tracking, i.e., the human observers are more sensitive to degradation than to improvement in picture quality. We adopt the implementation in [24] which is a lowpass function of the frame-level scores

$$f_n^{'} = \begin{cases} f_{n-1}^{'} + a_- \triangle_n, & \text{if} \quad \triangle_n \leq 0 \\ f_{n-1}^{'} + a_+ \triangle_n, & \text{if} \quad \triangle_n > 0 \end{cases} \tag{16}$$

where $\triangle_n = f_n - f_{n-1}^{'}$. The value difference between a_- and a_+ embodies the asymmetric tracking human behavior. In [24], the values of a_- and a_+ are derived by training ($a_- = 0.04$ and $a_+ = 0.5$). We re-tune the two parameters on our training set, and get similar results: $a_- = 0.075$, $a_+ = 0.431$. This parameter setting is used in our VQM. Finally, the overall sequence-level quality score s is given by averaging the intermediate results f_n', $n \in \{1, ..., N\}$

$$s = \frac{1}{N} \sum_{n=1}^{N} f_n^{'}. \tag{17}$$

IV. Experiments

A. Subjective Quality Video Databases and Performance Evaluation Criteria

In this section, two subjective video databases are used for performance evaluation, i.e., the LIVE and IVP subjective quality video databases. A subjective quality video database provides each of its distorted video a subjective score, which was obtained through subjective viewing tests. To evaluate predictive performance of a VQM, these subjective scores can be used as the ground truths to be compared against the metric's outputs. LIVE video database [45]–[47] consists of 150 $768 \times 432p$ distorted videos generated from ten reference videos of natural scenes with frame rates of either 25 or 50 f/s. Four practical distortion types for standard-definition videos are involved: H.264 compression, MPEG-2 compression, transmission errors over IP networks, and transmission errors over wireless networks. IVP video database [48] is developed by the author and his associates. It consists of 128 $1920 \times 1088p$ distorted videos with frame rate of 25 f/s. The reference videos contain both natural scenes and animations. Four practical distortions for high-definition videos are involved: H.264 compression, MPEG-2 compression, Dirac wavelet compression [49], and transmission errors over IP networks. Both databases pack their videos into YUV 4:2:0, and the duration of each distorted video sequence is around 10 s. The reference and distorted videos are well spatially

and temporally registered, and there is no need for luminance-chrominance alignment [50].

As required by some performance measures, such as the linear correlation coefficient, it is necessary to nonlinearly map each metric output (objective score) x to $Q(x)$, so that $Q(x)$ and the subjective scores approximately exhibit a linear relationship. We adopt the following nonlinear mapping function [17] for all VQMs in the experiments:

$$Q(x) = \beta_1 \times \left(0.5 - \frac{1}{1+\exp(\beta_2 \times (x-\beta_3))}\right) + \beta_4 \times x + \beta_5. \tag{18}$$

The fitting parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ are determined by minimizing the sum of squared differences between the mapped objective scores $Q(x)$ and the subjective scores. As to be introduced below, we divide the LIVE video database into a training set and a test set. In the following experiments, the fitting parameters are determined on the training set and are used to evaluate the predictive performances of the VQMs on the test set. As for the IVP video database, due to the limited number of test sequences, we do not divide it into a training set for determining the fitting parameters and a test set for the evaluation. On the other hand, the fitting parameters are chosen and tested on the entire IVP database.

The mapped scores $Q(x)$ and the subjective scores serve as inputs to three performance measures: the linear correlation coefficient (LCC), the root mean squared error (RMSE), the Spearman rank-order correlation coefficients (SROCC). For formulation and detailed comparison between these performance measures, please refer to [1] and [51]. In general, higher LCC indicates stronger correlation between objective and subjective scores, hence, better predictive performance; RMSE measures the predictive errors, so the smaller the RMSE the better the predictive performance. SROCC measures correlation between ranks of the objective and subjective scores, instead of their magnitudes as the LCC does. Therefore, it is immune to a failed monotonous nonlinear mapping.

B. Parameterization

There are four parameters to be determined, i.e., m_t, w, a_-, and a_+. We select 60 videos[4] from LIVE video database to train these parameters. The objective is to maximize the SROCC of the resultant VQM on this training set. In the training process, m_t is changed from 0 to 0.9 in step of 0.1. For each m_t, the best w is found by a global optimization algorithm, i.e., the genetic algorithm [52], with $a_- = 1$ and $a_+ = 1$ (i.e., $f_n^{'} = f_n$). Comparing all the $\{m_t, w\}$ pairs, $\{m_t = 0.4, w = 2.47 \times 10^3\}$ that maximizes the SROCC is chosen as the final parameters setting. Fig. 5 shows the m_t training results. When $m_t = 0.4$, SROCC reaches the peak. Further increase of m_t will degrade the predictive performance, since the temporal masking effect may be overestimated. After m_t and w are fixed, a_- and a_+ are found also by the genetic algorithm. Table I shows how the processing modules introduced in Section III accumulatively improve the predictive performance of the proposed VQM on both the training and test sets of LIVE video database, and the full set of IVP video database. It can be observed that by simulating the HVS processing using motion-based models as introduced in Section III, substantial improvements can be achieved in comparison to assessing quality directly after separating the impairments.[5] It is worth noting that similarly the spatial masking factor m_s can be tuned to further improve the predictive performance of the proposed VQM on the training set. However, since the more the tuning parameters, the more likely the algorithm over-fits the training data, the spatial masking factor m_s is simply set to 1 as in our previous work [1].

[4]The 60 distorted videos are generated from four reference sequences, i.e., *Tractor*, *Sunflower*, *Mobile and Calendar*, *Park Run*, randomly chosen from the LIVE video database.

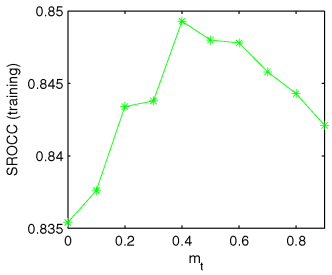


Fig. 5. Tuning results of the parameter m_t.

TABLE I
ACCUMULATIVE PERFORMANCE IMPROVEMENT (EVALUATED BY SROCC) OF THE PROPOSED VQM

	LIVE (Training)	LIVE (Test)	IVP
DCP	0.7911	0.6639	0.4478
DCP+CSF	0.8157	0.7527	0.7462
DCP+CSF+SM	0.8354	0.7832	0.7949
DCP+CSF+SM+TM	0.8493	0.8117	0.8059
DCP+CSF+SM+TM+AT	0.8590	0.8284	0.8392
CSF+SM+TM+AT	0.7712	0.6914	0.7901

DCP: decoupling, CSF: contrast sensitivity function, SM: spatial masking, TM: temporal masking, AT: asymmetric tracking.

C. With/Without Decoupling

To verify the benefits from the separation of detail losses and additive impairments, we develop another metric which uses the same HVS processing modules as those of the proposed VQM but treats the distortions integrally (i.e., without decoupling). To be more specific, the distortions, i.e., the differences between the original and distorted videos, are represented into the wavelet domain. Their intensities are modulated by the CSF, spatial masking,[6] and temporal masking models, as introduced in Section III, in order to simulate the HVS processing of the distortions. Equation (13), which is similar to the classical Minkowski summation typically used for information pooling, is applied in the spatial pooling. The approach discussed in Section III-G is employed as the temporal pooling process. The parameters (i.e., m_t, a_+, a_-) are tuned on the same training set. Its predictive performance

[5]In this case, the weighting factor w is also determined by using the training set of LIVE video database. Temporal pooling is implemented simply by averaging the frame-level quality scores.

[6]In this case, the original frames serve as the spatial maskers.

is shown in the last row of Table I. It can be observed that in comparison to the proposed VQM, the performance of this VQM without decoupling degrades significantly. It may serve as evidence that decoupling detail losses and additive impairments is indeed beneficial to video quality assessment.

D. Overall Predictive Performance

There are seven visual quality metrics tested in the experiments, i.e., PSNR, FSIM [53], IWSSIM [54], VSSIM [55], VQ model [5], MOVIE [19], and the proposed metric. Most of the codes were downloaded from the authors' websites, except for PSNR and VSSIM which were implemented by us. FSIM and IWSSIM are the representatives of the state-of-the-art image quality metrics. They are extended to video quality metrics simply by averaging the frame-level quality scores. FSIM assesses image quality by using two low-level features, i.e., the phase congruency and the gradient magnitude. IWSSIM is an improved version of image quality metric SSIM [14], where mutual information between the reference and distorted images is used to weight the SSIM index. VSSIM, VQ model, and MOVIE are popular video quality metrics. VSSIM is an extension of SSIM to video quality assessment, mainly by considering two observations that: 1) dark regions usually do not attract eye fixations; and 2) SSIM performs less stable when very large global motion occurs. Therefore, VSSIM assigns smaller weights to the dark regions and frames with large global motion. VQ model is a well-known reduced-reference video quality metric, which has been standardized in America (ANSI T1.801.03-2003) and was recommended by the ITU [56] due to its good performance in the VQEG Phase II validation tests [57]. It extracts statistical features (e.g., mean, standard deviation of the spatial luminance gradients) from 3-D cubes (e.g., 8 pixel $\times$ 8 lines $\times$ 0.2 s) for quality comparison. MOVIE is the state-of-the-art video quality metric which exploits motion information to weight distortions in 105 spatio-temporal frequency subbands, as introduced in Section II.

As mentioned in Section IV-B, 60 videos of the LIVE video database are used for the parameter training. The seven visual quality metrics are tested on the remaining 90 videos of LIVE and the full set of IVP (128 videos). The scatter plots of the proposed metric on the test sets are shown in Fig. 6. Each dot represents a test video. The vertical axis denotes the subjective ratings, and the horizontal axis denotes the nonlinearly mapped metric's outputs. Subjectively, if the dots scatter closely around the dashed line, then it means that the predictions of the metric and the subjective ratings of the human observers have a strong correlation. For illustration, Fig. 6 also shows the scatter plots of PSNR on both databases. By comparison, it can be observed that the scatter plots of the proposed metric demonstrate a stronger correlation between the metric's outputs and the subjective ratings.

The objective performance comparison is shown in Table II. On both video databases the proposed video quality metric achieves the state-of-the-art predictive performance in terms of the three performance evaluation criteria, i.e., LCC, SROCC, and RMSE. PSNR demonstrates poor performance on LIVE but quite good performance on IVP. The inconsistent performance of PSNR across the two databases may be due to

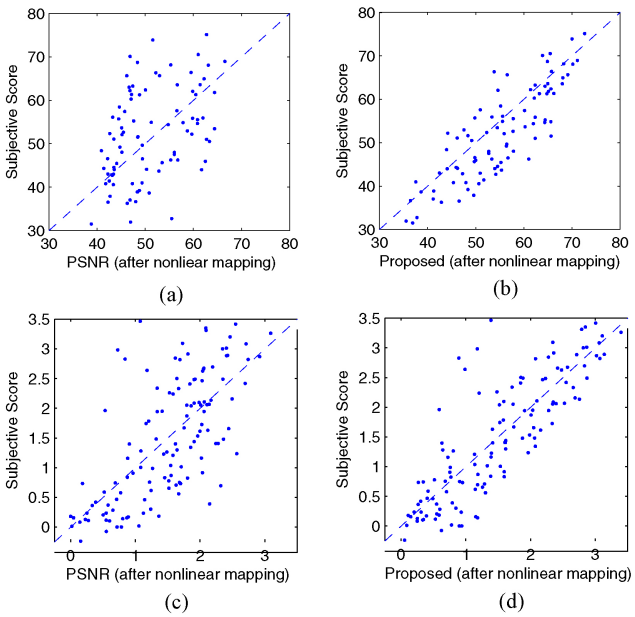


Fig. 6. Comparison between scatter plots of PSNR and the proposed VQM on test set of LIVE video database (90 videos) and full set of IVP video database (128 videos). (a) PSNR (LIVE test set). (b) Proposed (LIVE test set). (c) PSNR (IVP full set). (d) Proposed (IVP full set).

the fact that IVP contains a larger portion of videos with coding artifacts which is in favor of PSNR since it is well acknowledged that PSNR is better at handling coding artifacts than the transmission artifacts. Both FSIM and IWSSIM show relatively good performances on both databases, considering the fact that they measure spatial distortions only and do not exploit any temporal HVS characteristics for video quality assessment. Hence, they can serve as good start points for developing more advanced video quality metrics. The predictive performances of VSSIM on both databases are surprisingly low, probably because that SSIM is not effective enough as a spatial quality index and it does not make the best use of the motion information to simulate the HVS perception. As a practical reduced-reference video quality metric, VQ model performs relatively well, especially on the LIVE video database. MOVIE is the second best performer on LIVE, but due to its great demand for computing resources,[7] we cannot present its performance on IVP video database. It should be noted that none of the visual quality metrics in our experiments particularly takes into account how the video resolution affects quality perception. However, the proposed method takes the ratio of viewing distance to picture height (d/h) as input for CSF calculation, as shown in (8) and (11). For LIVE and IVP databases, this ratio is set to 6 and 3, respectively. This additional information may have favored the proposed metric in the performance comparison.

Table II also lists the results of the F-test. F-test assesses the statistical significance of the performance difference between two algorithms. In this paper, F-test is conducted on the prediction residuals of the quality metrics, i.e., the differences between the metric's outputs (after nonlinear mapping) and the subjective scores. The prediction residuals are supposed

[7]The code of MOVIE runs out of memory when processing HD videos.

TABLE II

PERFORMANCE COMPARISON BETWEEN SEVEN VISUAL QUALITY METRICS ON TWO SUBJECTIVE QUALITY VIDEO DATABASES, I.E., LIVE TEST SET (90 VIDEOS) AND IVP FULL SET (128 VIDEOS)

		PSNR	FSIM	IWSSIM	VSSIM	VQM	MOVIE	Proposed
LIVE (test set) $F_{\text{critical}} = 1.4199$	LCC	0.459	0.690	0.721	0.433	0.718	0.783	0.846
	SROCC	0.455	0.689	0.713	0.448	0.737	0.758	0.828
	RMSE	9.999	8.240	7.837	10.68	7.932	7.128	6.432
	Variance	99.46	63.69	60.49	113.0	61.44	48.23	34.83
	Kurtosis	2.36	2.55	2.52	2.24	3.07	2.99	2.56
	F-test	1	1	1	1	1	=	–
IVP (full set) $F_{\text{critical}} = 1.3404$	LCC	0.687	0.707	0.640	0.546	0.672	–	0.837
	SROCC	0.694	0.717	0.640	0.565	0.685	–	0.839
	RMSE	0.759	0.739	0.803	0.875	0.767	–	0.572
	Variance	0.580	0.550	0.649	0.771	0.591	–	0.330
	Kurtosis	3.50	4.87	4.54	3.20	4.72	–	5.03
	F-test	1	1	1	1	1	–	–

In the entries of F-test, symbols "1" and "=" respectively mean that the proposed metric is significantly better than or equivalent to the metric indicated by the first row of the table at 95% confidence level.

TABLE III

PERFORMANCE COMPARISON ON TWO DISTORTION SUBSETS (CODING: H.264+MPEG2, AND TRANS.: IP+WIRELESS) AND FIVE INDIVIDUAL DISTORTION TYPES (H.264, MPEG2, IP, DIRAC, WIRELESS), USING SROCC AS THE PERFORMANCE MEASURE

		PSNR	FSIM	IWSSIM	VSSIM	VQ Model	MOVIE	Proposed
Coding	LIVE(test)	0.257	0.656	0.649	0.467	*0.752*	*0.752*	**0.821**
	IVP	0.807	*0.878*	*0.857*	0.802	**0.904**	–	*0.896*
Trans.	LIVE(test)	0.507	*0.758*	*0.742*	0.429	*0.751*	*0.757*	**0.841**
H.264	LIVE(test)	0.225	0.650	0.662	0.461	0.737	*0.749*	**0.850**
	IVP	*0.782*	*0.851*	*0.802*	0.732	**0.864**	–	*0.842*
MPEG2	LIVE(test)	0.322	0.621	0.640	0.461	***0.865***	*0.808*	0.822
	IVP	0.741	*0.775*	0.733	*0.781*	**0.846**	–	*0.773*
IP	LIVE(test)	0.313	*0.701*	*0.721*	0.296	*0.777*	*0.647*	**0.806**
	IVP	*0.679*	0.551	0.390	*0.536*	*0.556*	–	**0.730**
Dirac	IVP	*0.846*	*0.894*	*0.880*	0.882	*0.912*	–	**0.923**
Wireless	LIVE(test)	*0.645*	*0.758*	*0.685*	0.455	*0.763*	**0.824**	*0.777*

The bold entries denote the best performer in terms of SROCC. The underlined entries denote the statistically best performer (with 95% confidence). The italic entries indicate metrics that are statistically indistinguishable from the underlined ones.

to be Gaussian, and smaller residual variance implies more accurate prediction. For example, to compare two metrics A and B, their residual variances V_A and V_B are calculated. Let F denote the ratio of V_A to V_B ($V_A > V_B$). If F is larger than F_{critical} which is calculated based on the number of residuals and a given confidence level, then the difference between the two metrics is considered to be significant at the specified confidence level. Table II lists the residual variances of each metric on the two databases, and also the F_{critical} values at 95% confidence level. The kurtosis values of the prediction residuals are provided in Table II as a measure of the Gaussianity: typically if the residuals have a kurtosis between 2 and 4 then they are taken to be Gaussian. In the entries of the F-test, symbol "1" or "=" denotes that the proposed method is statistically better than or equivalent to its competitor with 95% confidence. It is evident that on LIVE and IVP subjective quality video databases the proposed video quality metric outperforms most of its competitors statistically.

E. Performance on Individual Distortion Types

Table III shows the predictive performances of the VQMs on two distortion subsets (i.e., coding: H.264+MPEG2, and transmission error: IP+wireless) and five individual distortion types (i.e., H.264, MPEG2, IP, Dirac, wireless). For easier comparison, only the SROCC values are listed. SROCC is chosen because it is suitable for measuring a small number of data points and its value will not be affected by an unsuccessful monotonic nonlinear mapping. In Table III, the bold entries denote the best performer in terms of SROCC. The underlined entries denote the statistically best performer (with 95% confidence), and the italic entries indicate metrics that are statistically indistinguishable from the underlined ones. It can be observed from Table III that the proposed VQM demonstrates quite good performance. It achieves the highest SROCC on 6 of the 11 video sets. It is the statistically best performer on 7 of them, and in all the other cases, its performance is statistically equivalent to the best performer. Due to the limited number of videos in each subset, we cannot draw a solid conclusion on the best VQM regarding individual distortion types. However, it can be concluded in general that VQ model and the proposed method deliver relatively better results. Although its overall performance is intermediate as illustrated in Table II, VQ model performs quite well for individual distortion types, especially MPEG2 compression. It may be due to the fact that the parameters of VQ model are tuned by using many MPEG2 compressed videos [60].

TABLE IV
CROSS-DISTORTION PERFORMANCE EVALUATION (SROCC) OF THE PROPOSED VQM ON LIVE VIDEO DATABASE

Training Set \ Test Set	H.264	MPEG2	IP	Wireless
H.264	**0.926**	0.830	0.764	0.789
MPEG2	0.917	**0.893**	0.758	0.775
IP	0.852	0.785	**0.842**	0.771
Wireless	0.909	0.790	0.778	**0.800**

The bold entries denote the best performances for each distortion type (i.e., training and testing on the same data set).

TABLE V
CROSS-DISTORTION PERFORMANCE EVALUATION (SROCC) OF THE PROPOSED VQM ON IVP VIDEO DATABASE

Training Set \ Test Set	H.264	MPEG2	IP	Wireless
H.264	**0.854**	0.799	0.926	0.741
MPEG2	0.820	**0.849**	0.903	0.752
Dirac	0.829	0.798	**0.931**	0.735
IP	0.823	0.818	0.908	**0.790**

F. Cross-Distortion Performance Evaluation

Tables IV and V show the experimental results of cross-distortion performance evaluation of the proposed VQM on LIVE and IVP video databases, respectively. As in Table III, only the SROCC values are listed, which will not be changed by any monotonic nonlinear fitting. The first column indicates the training set, while the first row indicates the test set. As introduced in Section IV-B, four parameters are tuned by using the training set. The bolded diagonal entries denote the best performances (training and testing on the same data set) of the proposed VQM for each distortion type.

Observing the experimental results, it seems that given the optimal parameter values for each distortion type the proposed VQM demonstrates a large performance variance across different distortion types. In general, the proposed VQM tends to be better at evaluating coding artifacts (H.264, MPEG2, Dirac) than transmission artifacts (IP, wireless) on each database. Transmission artifacts typically occur in local regions which often attract viewer's attention. They cause other distortions less noticeable. By incorporating visual saliency model into our VQM, the above-mentioned phenomena can be taken into account and the transmission artifacts may be better handled. From Table V, it can be observed that no matter what the training set is, the predictive performance of the proposed VQM is always good for Dirac wavelet coding artifacts. The reason may be that Dirac wavelet coding only employs the intracoding mode and according to our observation the visual quality across sequential frames is quite smooth. These properties of the Dirac coded videos make the quality prediction much easier.

V. CONCLUSION

In this paper, we proposed a novel full-reference video quality metric. Based on our previous work [1], two distinct types of spatial distortions, i.e., detail losses and additive impairments, were decoupled and evaluated in the wavelet domain. Motion estimation was performed to derive motion vectors which then were used in the simulation of the HVS processing of the spatial distortions. The simulated HVS characteristics included the HVS contrast sensitivity modeled by Daly's eye movement spatio-velocity CSF, and both spatial and temporal visual masking implemented in an engineering manner. The quality impacts of the two types of spatial distortions were assessed using distinct equations and integrated simply by weighted summation. Ultimately, the frame-level quality scores were processed taking into account cognitive human behaviors and averaged to generate the sequence-level quality score depicting the perceptual visual quality of the whole video sequence. Rather than test on only one database as most previous studies did [19], [20], [30], [34], [36], [50], [55], [58], [59], two subjective quality video databases, i.e., LIVE and IVP, are used for performance comparison. It can be observed from the experimental results that on both databases the proposed full-reference video quality metric achieves the state-of-the-art performance in matching subjective ratings.

REFERENCES

[1] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[2] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Rec. BT.500-11, ITU, Geneva, Switzerland, 2002.

[3] *Subjective Assessment Methods for Image Quality in High-Definition Television*, ITU-R Rec. BT.710-4, ITU, Geneva, Switzerland, 1998.

[4] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T Rec. P.910, ITU, Geneva, Switzerland, 2008.

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

[6] K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra- and interframe reduced-reference features," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3229–3232.

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

[8] M. Narwaria and W. S. Lin, "Objective image quality assessment based on support vector regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 515–519, Mar. 2010.

[9] S. Daly and A. B. Watson, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993, pp. 179–205.

[10] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *Proc. 24th Soc. Inform. Display Dig. Tech. Papers*, 1993, pp. 946–949.

[11] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE ICIP*, vol. 2. Nov. 1994, pp. 982–986.

[12] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. Bovik, Ed. New York: Academic Press, 2000.

[13] H. R. Sheikh, Z. Wang, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, E. B. Furht and O. Marqure, Eds. Boca Raton, FL: CRC Press, 2003, pp. 1041–1078.

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

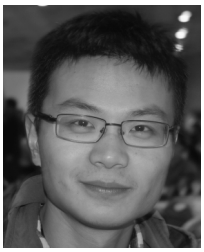
[15] C. L. Yang, W. R. Gao, and L. M. Po, "Discrete wavelet transform-based structural similarity for image quality assessment," in *Proc. IEEE ICIP*, Oct. 2008, pp. 377–380.

[16] M. Zhang and X. Q. Mou, "A psychovisual image quality metric based on multi-scale structure similarity," in *Proc. ICIP*, vols. 1–5. 2008, pp. 381–384.

[17] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[18] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *Proc. ICIP*, vols. 1–5. 2008, pp. 1200–1203.
[19] 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.
[20] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *Opt. Eng.*, vol. 42, no. 1, pp. 265–272, 2003.
[21] J. Lubin, "A human vision system model for objective picture quality measurements," in *Proc. Int. Broadcasting Conv.*, Sep. 1997, pp. 498–503.
[22] S. Winkler, "A perceptual distortion metric for digital color video," in *Proc. 4th Hum. Vis. Electron. Imaging*, vol. 3644. 1999, pp. 175–184.
[23] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, no. 1, pp. 20–29, 2001.
[24] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 260–273, Feb. 2006.
[25] F. X. J. Lukas, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. 30, no. 7, pp. 1679–1692, Jul. 1982.
[26] P. Lindh and C. J. van den Branden Lambrecht, "Efficient spatio-temporal decomposition for perceptual processing of video sequences," in *Proc. ICIP*, vol. III. Sep. 1996, pp. 331–334.
[27] C. H. Chou and C. W. Chen, "A perceptually optimized 3-D subband codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 2, pp. 143–156, Feb. 1996.
[28] J. M. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Opt. Soc. Amer. A Opt. Image Sci. Vis.*, vol. 11, no. 6, pp. 1710–1719, 1994.
[29] Z. Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.
[30] E. P. Ong, X. K. Yang, W. S. Lin, Z. K. Lu, S. S. Yao, X. Lin, S. Rahardja, and B. C. Seng, "Perceptual quality and objective quality measurements of compressed videos," *J. Vis. Commun. Image Representation*, vol. 17, no. 4, pp. 717–737, 2006.
[31] S. S. Yao, W. S. Lin, E. P. Ong, and Z. K. Li, "A wavelet-based visible distortion measure for video quality evaluation," in *Proc. ICIP*, vols. 1–7. 2006, pp. 2937–2940.
[32] W. Lin, "Gauging image and video quality in industrial applications," in *Advances of Computational Intelligence in Industrial Systems*, Y. Liu, *et al.*, Eds. Heidelberg, Germany: Springer-Verlag, 2008, pp. 117–137.
[33] W. S. Lin, "Computational models for just-noticeable difference," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2005.
[34] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A: Opt. Image Sci. Vis.*, vol. 24, no. 12, pp. B61–B69, 2007.
[35] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 266–279, Apr. 2009.
[36] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.
[37] K. T. Tan, M. Ghanbari, and D. E. Pearson, "An objective measurement tool for MPEG video quality," *Signal Process.*, vol. 70, no. 3, pp. 279–294, 1998.
[38] Y. Horita, T. Miyata, I. P. Gunawan, T. Murai, and M. Chanbari, "Evaluation model considering static-temporal quality degradation and human memory for SSCQE video quality," in *Proc. Vis. Commun. Image Process.*, vol. 5150. 2003, pp. 1601–1611.
[39] S. Daly and C. J. van den Branden Lambrecht, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision Models and Applications to Image and Video Processing*. Norwell, MA: Kluwer, 2001, pp. 179–200.
[40] S. Li, L. Ma, and K. N. Ngan, "Video quality assessment by decoupling additive impairments and detail losses," in *Proc. 3rd Int. Workshop Qual. Multimedia Experience*, 2011, pp. 90–95.
[41] Y. Q. Zhang and S. Zafar, "Motion-compensated wavelet transform coding for color video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, no. 3, pp. 285–296, Mar. 1992.
[42] D. H. Kelly, "Motion and vision II: Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Amer.*, vol. 69, no. 10, pp. 1340–1349, 1979.
[43] J. G. Robson, "Spatial and temporal contrast-sensitivity functions of visual system," *J. Opt. Soc. Amer.*, vol. 56, no. 8, pp. 1141–1142, 1966.
[44] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly, "Spatio-velocity CSF as a function of retinal velocity using unstabilized stimuli," *Proc. SPIE Conf. Human Vision Electron. Imag. XI*, vol. 6057, p. 5705, Jan. 2006.
[45] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. (2009). *LIVE Video Quality Database* [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html
[46] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[47] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," *Proc. SPIE*, vol. 7527, p. 75270H, Jan. 2010.
[48] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. (2011). *IVP Video Quality Database* [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/index.html
[49] BBC Research. (2008). *Dirac Video Coding Standard* [Online]. Available: http://diracvideo.org/specifications
[50] A. P. Hekstraa, J. G. Beerendsa, D. Ledermannb, F. E. de Caluwec, S. Kohlerb, R. H. Koenen, S. Rihs, M. Ehrsam, and D. Schlaussb, "PVQM: A perceptual video quality measure," *Signal Process.-Image Commun.*, vol. 17, no. 10, pp. 781–798, 2002.
[51] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[52] A. Chipperfield, P. Fleming, H. Pohlheim, and C. Fonseca. (1994). *Genetic Algorithm Toolbox* [Online]. Available: http://www.shef.ac.uk/acse/research/ecrg/gat.html
[53] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[54] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[55] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.
[56] *Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference*, ITU-R Rec. BT.1683, ITU, Geneva, Switzerland, 2004.
[57] Video Quality Expert Group (VQEG). (2003). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment II* [Online]. Available: http://www.vqeg.org
[58] W. S. Lin, D. Li, and X. Ping, "Visual distortion gauge based on discrimination of noticeable contrast changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 900–909, Jul. 2005.
[59] E. P. Ong, W. S. Lin, Z. K. Lu, S. Yao, and M. Etoh, "Visual distortion assessment with emphasis on spatially transitional regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 4, pp. 559–566, Apr. 2004.
[60] S. Wolf and M. Pinson. (2002, Jun.). *Video Quality Measurement Techniques, NTIA Rep. 02-392* [Online]. Available: www.its.bldrdoc.gov/n3/video/documents.htm

Songnan Li (S'08) received the B.S. and M.S. degrees from the Department of Computer Science, Harbin Institute of Technology, Harbin, China, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

His current research interests include visual quality assessment, video deinterlacing, video compression, code optimization.



Lin Ma (S'09) received the B.E. and M.E. degrees, both in computer science, from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Shatin, Hong Kong.

He was a Research Intern with Microsoft Research Asia, Beijing, China, from October 2007 to March 2008. He was a Research Assistant with the Department of Electronic Engineering, CUHK, from November 2008 to July 2009. He was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore, from July 2011 to September 2011. His current research interests include image/video quality assessment, super-resolution, restoration, and compression.

Mr. Ma received the Best Paper Award in the Pacific-Rim Conference on Multimedia in 2008. He was awarded the Microsoft Research Asia Fellowship in 2011.

King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from Loughborough University, Loughborough, U.K.

He is currently a Chair Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. He was previously a Full Professor with Nanyang Technological University, Singapore, and with the University of Western Australia, Crawley, Australia. He holds honorary and visiting professorships with numerous universities in China, Australia, and South East Asia. He has published extensively including three authored books, six edited volumes, over 300 refereed technical papers, and has edited nine special issues in journals. He holds ten patents in the areas of image/video coding and communications.

Dr. Ngan served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal on Visual Communications and Image Representation*, the *EURASIP Journal of Signal Processing: Image Communication*, and the *Journal of Applied Signal Processing*. He has chaired a number of prestigious international conferences on video signal processing and communications, and has served on the advisory and technical committees of numerous professional organizations. He co-chaired the IEEE International Conference on Image Processing, Hong Kong, in September 2010. He is a Fellow of IET (U.K.) and IEAust (Australia), and was an IEEE Distinguished Lecturer from 2006 to 2007.