Novel Spatio-Temporal Structural Information Based Video Quality Metric

Yue Wang, Tingting Jiang, Siwei Ma, *Member, IEEE,* and Wen Gao, *Fellow, IEEE*

***Abstract*—Video quality assessment (VQA) is very important for many video processing applications, e.g., compression, archiving, restoration, and enhancement. An ideal video quality metric should achieve consistency between video distortion prediction and psychological perception of human visual system. Different from the quality assessment of single images, motion information and temporal distortion should be carefully considered for VQA. Most of previous VQA algorithms deal with the motion information through two ways: either incorporating motion characteristics into a temporal weighting scheme to account for their affects on the spatial distortion, or modeling the temporal distortion and spatial distortion independently. Optical flows need to be estimated in the two ways. In this paper, we propose a different methodology to deal with the motion information. Instead of explicitly calculating the optical flow and independently modeling the temporal distortion, both the spatial edge features and temporal motion characteristics are accounted for by some structural features in the localized space-time regions. We propose to represent the structural information by two descriptors extracted from the 3-D structure tensors, which are the largest eigenvalue as well as its corresponding eigenvector. Experimental results on LIVE database and VQEG FR-TV Phase-I database show that the proposed VQA metric is competitive with state-of-the-art VQA metrics, while keeping relatively low computing complexity.**

***Index Terms*—3-D structure tensor, human visual system (HVS), video quality assessment (VQA).**

I. Introduction

OWING TO THE RAPID development of digital media applications, digital video resources have been explosively increasing in the past decades. Quality assessment of these media resources is exceedingly important in the systems of digital video acquisition, compression, transmission, and storage. The most straightforward way to evaluate the quality of a video is to use the quality scales directly rated by human observers. However, such subjective evaluations are quite time-consuming and expensive, and could not be applied in real-time scenarios and automatic systems. Therefore, there has been an increasing demand for objective quality criteria and metrics that are in agreement with the HVS' judgment.

Objective video quality metrics can be generally classified into three distinct categories according to the availability of the reference video signals: full-reference (FR) metrics, reduced-reference metrics, and no-reference metrics. The proposed method as well as all the previous works investigated in this paper belongs to the first category, in which both the reference and distorted videos could be accessed by the algorithms. In general, applications of FR video quality assessment (VQA) metrics include but are not limited to [1]:

1) codec evaluation, specification, and acceptance testing;
2) in-service quality monitoring at the source;
3) remote destination quality monitoring when a copy of the source is available;
4) quality measurement of a storage or transmission system that utilizes video compression and decompression techniques.

Real-time processing capability of a FR VQA metric is quite important for in-service quality monitoring at source end and when it is used as the distortion metric in a practical rate-distortion optimized video encoder. For other applications, low complexity is also a desirable advantage.

The pixel-based FR metrics such as mean-squared error and related peak signal-to-noise ratio (PSNR) have been the dominant quantitative performance metrics in the field of signal processing for several decades since they are simple to calculate and have clear physical meanings in terms of Shannon information theory. However, it has been well acknowledged that these pixel-based signal fidelity metrics do not always correlate well with the HVS' perception [2], [3]. For the perception of the HVS, the visual information obtained from natural videos is not reflected in the individual pixels, but in some high-order statistics of the pixels in both the spatial domain and the temporal domain, which represent the object's structural features and motion characteristics. In an effort to take into account the features of the HVS, many FR approaches and methodologies for VQA have been proposed in the past few years.

One obvious and simple way to perform VQA is to implement quality evaluation on each individual frame and then summate all the frame scores to obtain a composite score. Many image quality metrics have been directly extended to VQA metrics using a frame-by-frame approach

Manuscript received July 8, 2011; revised October 5, 2011; accepted November 3, 2011. Date of publication February 3, 2012; date of current version June 28, 2012. This work was supported in part by the National Basic Research Program of China (973 Program, 2009CB320903) and the National Science Foundation, under Grants 61121002, 60833013, and 60803068. This paper was recommended by Associate Editor M. Tagliasacchi.

Y. Wang is with the Graduate University of the Chinese Academy of Sciences, Beijing 100080, China (e-mail: wangyue@jdl.ac.cn).

T. Jiang, S. Ma, and W. Gao are with the National Engineering Laboratory for Video Technology and Key Laboratory of Machine Perception (MoE), School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: ttjiang@pku.edu.cn; swma@pku.edu.cn; wgao@pku.edu.cn).

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.2012.2186745

and a proper temporal weighting scheme. For example, a quite representative quality metric of image quality assessment (IQA) is the SSIM index proposed by Wang *et al*. [4], which evaluates image quality by using some low level structural information such as mean, variance, and covariance of intensity values of pixels in local patches. To produce a video quality score based on the individual frame quality scores, a motion-weighting model [5] is proposed to account for the fact that the accuracy of visual perception is significantly reduced when the speed of motion is large, and in [6], an alternate weighting scheme based on human perception of motion information is utilized. Although motion information is more or less explored in these temporal weighting schemes, however, temporal distortions were not yet taken into account [8].

In recent years, temporal distortion has drawn more and more attention from the VQA researchers. Video quality metric (VQM) proposed by NTIA [7] is a popular VQA metric that was included in the Recommendation ITU-T J.144 [1] as a normative FR VQA model. This metric extracts seven features from spatio-temporal blocks to compute the video distortion. Frame differences are embedded into one feature to account for the interaction between motion and spatial distortion. In [9], temporal distortion is defined as temporal evolution of the spatial distortion in a spatio-temporal "tube," since the perception of spatial distortions over time can be largely modified by their temporal changes. Seshadrinathan *et al.* proposed a motion-based video integrity evaluation (MOVIE) index in [8], where they defined the temporal distortion as the differences between the filter responses along computed motion trajectories. A similar framework is proposed in [10], which calculates the temporal distortion as the SSIM index between motion compensated video patches. With the block-based motion compensation, this method is quite computationally efficient.

The methodologies of previous works that deal with motion information could be roughly summarized in two categories, either incorporating motion characteristics into the weighting factors to account for their effects on the spatial distortion, or modeling the temporal distortion and spatial distortion independently. In many cases, it is difficult to strictly distinguish the spatial distortion from the temporal distortion. As the natural videos could be regarded as pixels arranged along two spatial and one temporal dimensions, both spatial structure features and temporal motion characteristics are reflected in some high-order statistics in the localized spacetime regions. The localized distortions would impair the local statistics in both the spatial domain and the temporal domain. For example, blurring effect would not only obscure the HVS' perception to the edge contours, but also introduce a false perception of the localized motion. Other spatial artifacts such as blocky and ringing artifacts would also impair the HVS' perception to the motion continuity.

In this paper, we propose a new methodology to deal with the motion information. Instead of explicitly calculating the optical flow and independently modeling distortions in temporal and spatial domains, we extract structural descriptors from the localized spacetime regions to account for spatial and temporal distortions simultaneously. The largest eigenvalue and its corresponding eigenvector of the 3-D structure tensors are used as the descriptors because they can well represent spatio-temporal structural features of the localized spacetime regions. Spatial structure features as well as motion information are implicitly represented in the descriptors. Furthermore, a double saliency detection mechanism is incorporated into the metric, making it efficient in prediction consistency and computing complexity. This metric has a clear physical meaning in accordance with visual perceptions of the HVS and is quite computationally efficient. Experimental results demonstrate that the proposed metric's performance is competitive with other state-of-the-art VQA metrics.

The remainder of this paper is organized as follows. In Section II, we focus on the details of the proposed video quality metric. Simulation results are presented in Section III. Finally, Section IV concludes this paper.

II. Proposed Method

Structural information based image quality metrics [4], [11]–[13] have been widely studied in recent years. The motivation of these methods is that HVS is highly adapted to extract the structural information from the visual scene. Natural images are not random collections of pixels, but have strong statistical dependencies between the pixels. HVS understands natural images based on some low level structural features, which implicitly present in the relationship between the pixels. Accordingly, perceptible degradations of images correspond to distortions of the structural features. Therefore, the structural information based IQA metrics are devoted to find the representative structural features which are in consistence with HVS.

In this paper, we extend the insights of structural similarity to spatio-temporal case. It is widely accepted that the basic primitives and dominant features of natural images are the edges [14]. Along the temporal axis, displacement of these spatial primitives gives HVS the perception of motion. If we regard the video as a pixel volume, the local spacetime region would exhibit highly spatio-temporally structured characteristics. As illustrated in Fig. 1, the edge contour of a moving object would stretch out a plane along its motion trajectory in the spatio-temporal space. As a result, the variation of the gray value in the localized spacetime region would be oriented to a certain orientation which we refer to as the primary direction. Both the object edge $\vec{e}$ and its motion trajectory $\vec{v}$ would lie in the plane which is orthogonal to this primary direction $\vec{p}$. Perceptible degradations of edge and motion would alter the energy distribution in the spacetime region.

There are many mathematical tools that could be used to describe these spatio-temporal structural features, such as steerable filters [15], least-square estimation [16], and anisotropic diffusion [17]. In this paper, we leverage a powerful tool: the 3-D structure tensor [18], in which the spatio-temporal oriented structural information is implicitly embedded into a local gradient based matrix.

A. Introduction of 3-D Structure Tensor

The structure tensor, since first introduced by Harris [19] in the task of corner and edge detection, has proven its power in

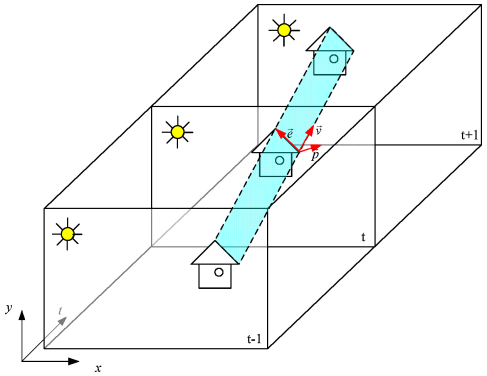


Fig. 1. Illustration of the local spatio-temporal structure features.

many applications like texture analysis [20], diffusion filtering [21], optical flow estimation and analysis [22], [23], and video segmentation [24]. The structure tensor is a matrix derived from the local gradients, whose eigenvectors and eigenvalues summarize the predominant directions of the energies in a specified neighborhood of a point, and the coherency of those directions. In particular, for the 3-D video data, the 3-D structure tensor at point p has the following mathematical form:

$$S(p) = \nabla I(p) \cdot \nabla I(p)^T = \begin{bmatrix} \sum_W I_x^2(p) & \sum_W I_x(p) \cdot I_y(p) & \sum_W I_x(p) \cdot I_t(p) \\ \sum_W I_x(p) \cdot I_y(p) & \sum_W I_y^2(p) & \sum_W I_y(p) \cdot I_t(p) \\ \sum_W I_x(p) \cdot I_t(p) & \sum_W I_y(p) \cdot I_t(p) & \sum_W I_t^2(p) \end{bmatrix} \quad (1)$$

where $\nabla = (\partial_x, \partial_y, \partial_t)$ denotes partial derivatives along x, y, and t directions, respectively, and W is a local integration window. The 3-D structure tensor is a 3×3 symmetric matrix which contains six independent components. The matrix has two important advantages for structure analysis [23]. First, the matrix representation of the gradients allows the integration of information from a local neighborhood without cancelation effects which would appear if gradients with opposite orientation were integrated directly. Second, the integration of local orientation in a window yields robustness against noise, and creates additional information such as coherence.

The localized spatio-temporal structural information is contained in the eigenvectors and eigenvalues of this matrix, which will be further analyzed in the next section.

B. Structural Descriptors Extracted From the 3-D Structure Tensor

The importance of this matrix stems from the fact that its eigenvalues and the corresponding eigenvectors summarize the distribution of the energy within the local window centered at p. After performing eigenvalue decomposition of the 3×3 matrix, the 3-D structure tensor can be expressed as follows:

$$S(p) = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \lambda_3 e_3 e_3^T \quad (2)$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ are the three eigenvalues sorted in descending order, and e_1, e_2, e_3 are their corresponding eigenvectors.

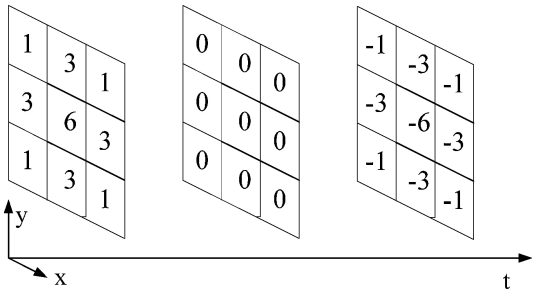


Fig. 2. Sobel kernel for t direction.

The eigenvectors represent the local orientations along which the local gray value variation is aligned, while their corresponding eigenvalues denote the variations along these directions. From the perspective of principal component analysis [25], the eigenvector e_1 corresponds to the primary direction in Fig. 1, which incurs the largest variance of the localized spacetime region, while the largest eigenvalue λ_1 reveals the degree of data variation along this direction. In another word, they represent the "strength" and the orientation of the 3-D structure. Accordingly, they are the most representative and efficient descriptors to account for structural distortion. As analyzed above, perceptible degradations of edge and motion correspond to structural distortion of the spacetime region, thus they could also be reflected in the alteration of these two descriptors.

Edge sharpness is usually measured by the gradient variation orthogonal to the edge [26], which is in consistence with the physical meaning of the eigenvalue. Blurring an edge would lead to decrease of the eigenvalue.

Another typical distortion introduced by block-based compression is the "blocky" artifact. This type of distortion introduces nonexisting edges in the spatial domain, obscuring the HVS' perception of the localized motion. The structural feature in the localized spacetime region would be largely changed by such an artifact but also reflected in the two descriptors.

As for the temporal distortion introduced by packet loss, which gives us the feeling of flickering, the abrupt change between adjacent frames would turn the direction of the eigenvector toward the temporal axis, and the corresponding eigenvalue altered as well.

C. Proposed Algorithm

The proposed VQA metric is performed only on the "Y" component of the video sequences. The first step of the algorithm is to collect gradient information. In this paper, we apply the 3-D Sobel kernels to calculate the local gradients. Fig. 2 shows the kernel for calculating the gradients along time direction. This is a $3 \times 3 \times 3$ matrix, which means for the pixels in current frame, we need two adjacent frames to calculate their gradients.

The kernels for x and y directions could be obtained by rotating this kernel by 90° along the y-axis and x-axis, respectively.

After the pixel gradients are obtained, we need to determine which pixels need to be processed. Recently, visual attention

has been widely investigated in VQA studies [27], [28]. Since human attention is not allocated equally to all regions in the visual field, but focused on certain regions known as salient regions [29], it is believed that distortion in these salient regions plays a crucial rule in the HVS' judgment on the overall quality of the video. Such an attention related mechanism is incorporated into our algorithm. In contrast with previous works which mostly perform saliency detection only on the original video, we detect salient pixels in both reference video and distorted video, and the final selected pixels for processing are the union of the salient pixels in original video and distorted video. This procedure is based on the consideration that the distortion process may introduce some salient artifacts in certain regions that did not draw the HVS' attention in the original video. However, these salient artifacts would greatly affect the HVS' judgment on the video quality. Fig. 3 illustrates the double saliency detection process. In the original video, the yellow pixels are the detected salient pixels. In the distorted video, the green pixels are the detected salient pixels. As a result, the pixels which require similarity analysis are the red pixels.

As to the saliency detection method, an efficient and fast algorithm is desired. Considering HVS is sensitive to edge distortion, ITU-T J.144 recommended a FR video metric, EPSNR, which applies thresholding to the spatial gradient magnitude to find edge pixels, and then calculate PSNR only on these edge pixels. In this paper, we consider that HVS is sensitive to edges, motion regions and abruptly emerged artifacts, where grayvalues of pixels usually change dramatically. Therefore, we judge a pixel is a salient pixel if its spatio-temporal gradient magnitude is above a certain threshold in either original video or distorted video. Discarding the nonsalient pixels not only improves the accuracy of quality evaluation, but also saves up much computing time, which is quite important for a real-time VQA metric.

When the salient pixels have been selected, we construct a pair of 3-D structure tensors for each salient pixel in both the reference video and the distorted video, and then perform eigenvalue decomposition on both. We utilize the widely used Jacobi method [30] in this paper. The largest eigenvalues and their corresponding eigenvectors are retained as the descriptors which are further used to calculate the quality score at this pixel according to the following formula:

$$m = \frac{2 \cdot l_r \cdot l_d}{l_r^2 + l_d^2} \times \cos < e_r, e_d > \tag{3}$$

where l_r and l_d denote the largest eigenvalues of the structure tensors in the reference video and distorted video, while e_r and e_d denote their corresponding eigenvectors.

The first term measures the similarity between the variances along their primary directions in the localized spacetime region, and the second term measures the divergence of their primary directions. Both terms and their product lie in the range of [0, 1]. This score indicates the degree of structural similarity between the corresponding localized spacetime regions at the same position, where a higher value indicates a better quality.

Finally, all of the salient pixel scores are averaged to give a final video quality index.

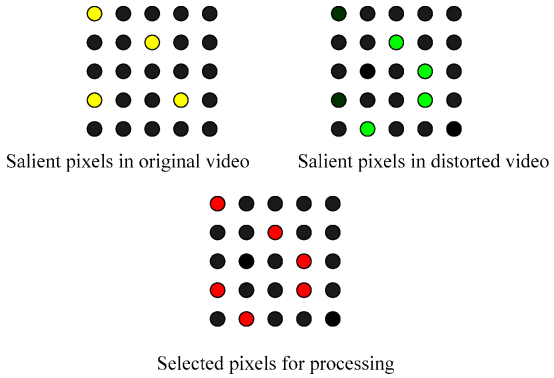


Fig. 3. Illustration for double saliency detection.

The specific procedure of the proposed VQA algorithm is given in Table I, and Fig. 4 illustrates the flow chart of the algorithm. Furthermore, we illustrate the performance of the proposed method by showing the saliency maps, eigenvalue maps and quality maps in Fig. 5. We select a pair of frames from one of the test sequences in the LIVE database. The corresponding video of (B) is generated by H.264 compression and network packet loss. To give a better visual effect, all the nonsalient pixels are dyed green. It is evident that the generated saliency maps of original video and distorted video are different. Some salient artifacts emerge in the distorted frame while these regions do not draw the HVS' attention in the original frame. These artifact regions have different responses in the eigenvalue maps and correspond to lower values in the final quality map.

III. Simulation Results

To evaluate the performance of the proposed VQA metric, we test it alongside other state-of-the-art VQA metrics on two publicly-accessible databases, namely the LIVE Video Quality Database [31], and the VQEG Phase I FR-TV test dataset [32]. The LIVE Video Quality Database consists of 10 reference videos with 15 distortions each, to give a total of 150 distorted videos, and the VQEG database contains 20 reference sequences and 16 distorted versions of each reference, for a total of 320 videos. Subjective scores (DMOS) were recorded for all test sequences in both datasets. The main difference between the two databases is in the types of distortions. VQEG Phase I dataset is largely comprised of compressed videos, while sequences in LIVE Video Quality Database are distorted by four different distortion processes—MPEG-2 compression, H.264 compression, and simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks [31]. As a result, the LIVE Video Quality Database contains more types of distortions, especially spatio-temporally localized distortions.

One thing that should be noticed is that there are two distortion types in the VQEG database (HRC 8 and 9), each containing two different subjective scores according to whether these sequences were viewed along with "high" or "low" quality videos [33]. As same as the test condition in [8], we used the scores assigned in the "low" quality regime as the subjective scores for these videos. Since most of the sequences

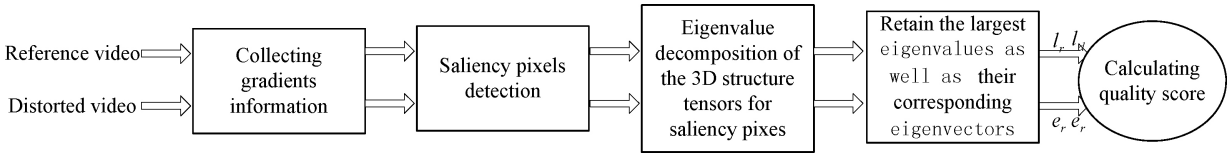


Fig. 4. Block diagram of the proposed algorithm.

TABLE I
PROCEDURE OF THE PROPOSED VQA ALGORITHM

Input:
The reference video and the distorted video
Output:
Quality score for the distorted video
Procedure:
1. Using the Sobel operator, calculates pixel gradients of the Y component along x, y and t directions in both reference video and distorted video, denote as gx_r, gy_r, gt_r and gx_d, gy_d, gt_d.
2. Calculate the saliency of each pixel. If $\sqrt{gx_r^2 + gy_r^2 + gt_r^2} > \varepsilon$ or $\sqrt{gx_d^2 + gy_d^2 + gt_d^2} > \varepsilon$, then this pixel is regarded as a salient pixel. ε is a predefined threshold.
3. For each salient pixel, construct a pair of 3D structure tensors according to (1) for both the reference video and the distorted video.
4. Perform eigenvalue decomposition on this pair of structure tensors. The largest eigenvalues l_r and l_d as well as their corresponding eigenvectors e_r and e_d are retained as structural descriptors for further calculation.
5. Calculate the quality score at this pixel according to (3).
6. Compute the quality score of a frame as the average of the quality scores over all the salient pixels in that frame.
7. Compute the quality score of a video as the average of the frame quality scores.

in VQEG database are in interlacing format, the proposed VQA algorithm is only applied on the top fields.

A. Performance Comparison

For comparison, these same sets of videos were evaluated by the following VQA metrics.

1) *PSNR*: the classic pixel-based VQA metric which is always used as baseline for performance evaluation of the VQA algorithms.
2) *VQM*: a widely used VQA metric proposed by NTIA [7], which was recommended by ITU J.144.
3) *SW-SSIM*: frame based SSIM with motion associated weighting [6].
4) *MOVIE*: the representative optical flow based VQA metric proposed in [8].
5) *MC-SSIM*: motion compensated SSIM which is proposed in [10].

Additionally, both of the databases are evaluated by the Tektronix PQA500 Picture Quality Analyzer, which is a leading video quality assessment product for industry application. Two indicators, namely PQR and DMOS exported by PQA500 are used for comparison.

Parameter configuration of the proposed metric is discussed in the next subsection.

TABLE II
PERFORMANCE COMPARISON ON THE LIVE DATABASE

Methods	Spearman CC	Pearson CC
MOVIE	0.786	0.810
VQM	0.702	0.723
SW-SSIM	0.585	0.596
MC-SSIM	0.679	0.698
PSNR	0.368	0.404
PQR (by PQA500)	0.695	0.712
DMOS (by PQA500)	0.695	0.711
Prop	0.779	0.778

As for performance criteria, Pearson correlation coefficient (CC) and Spearman rank order correlation coefficient are used as performance indicator. For the indicator of Pearson correlation coefficient, a nonlinear mapping between the objective scores and subjective quality ratings was used according to VQEG recommendations [33]. In this paper, the mapping function chosen for regression for each of the metrics was a 4-parameter logistic function

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp(-\frac{x-\tau_3}{\tau_4})} + \tau_2. \quad (4)$$

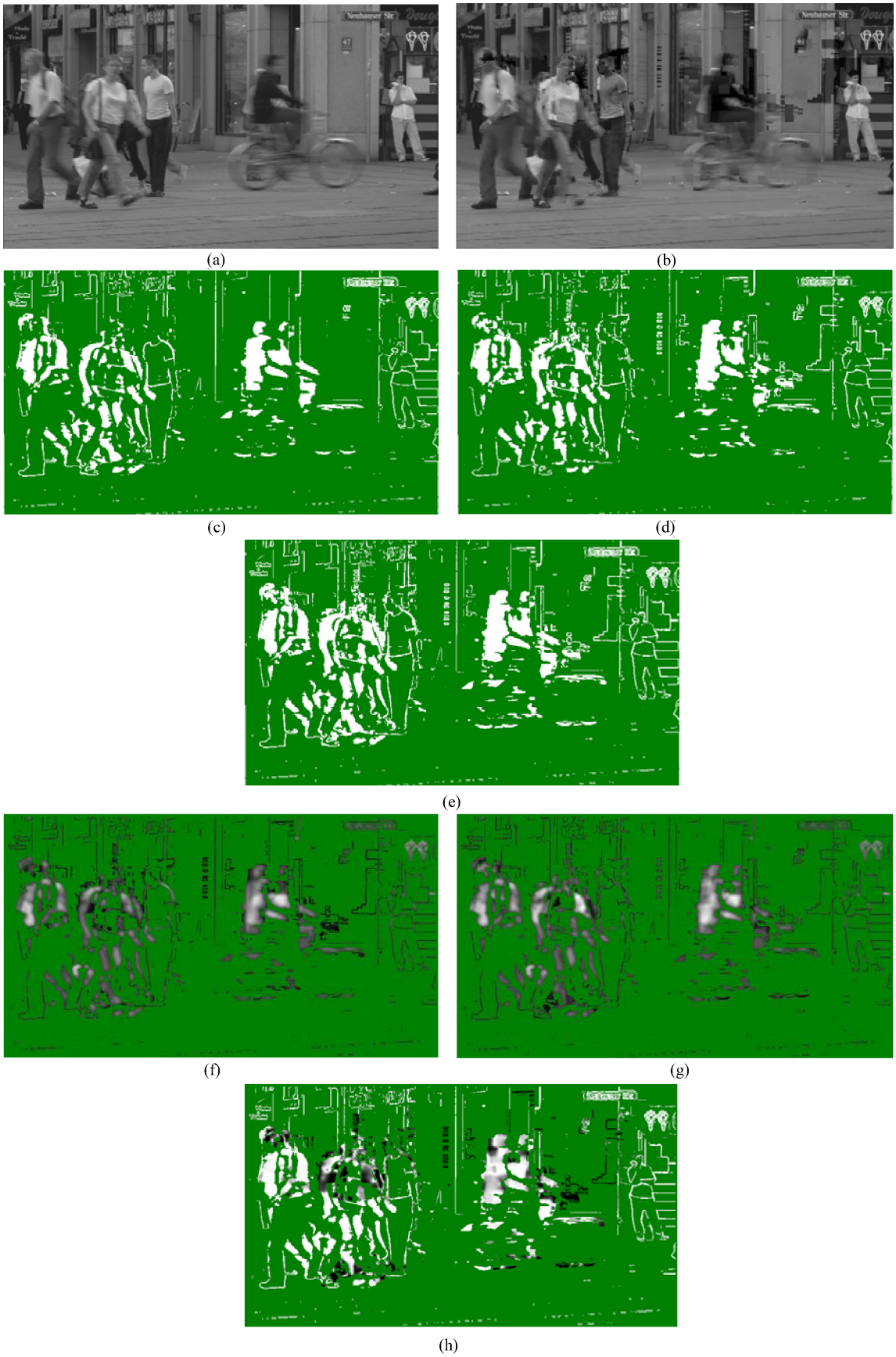


Fig. 5. Illustration of the performance of the proposed metric. (a) Frame from reference video. (b) Corresponding frame from distorted video. (c) Selected salient pixels of (a). (d) Selected salient pixels of (b). (e) Selected salient pixels by the double saliency detection. (f) Largest eigenvalues of the salient pixels in (a). (g) Largest eigenvalues of the salient pixels in (b). (h) Quality map of the salient pixels. Dark regions correspond to regions of poor quality.

TABLE III
PERFORMANCE COMPARISON ON THE VQEG DATABASE

Methods	Spearman CC	Pearson CC
MOVIE	0.833	0.821
VQM	0.781	0.782
SW-SSIM	0.837	0.810
MC-SSIM	0.833	0.833
PSNR	0.786	0.779
PQR (by PQA500)	0.814	0.814
DMOS (by PQA500)	0.815	0.814
Prop	0.832	0.822

TABLE IV
LCC SCORES OF VQA METRICS ON EACH KIND OF DISTORTION IN LIVE DATABASE

Methods	Wireless	IP	H.264	MPEG2
PSNR	0.4675	0.4108	0.4385	0.3856
SW-SSIM	0.5867	0.5587	0.7206	0.6270
VQM	0.7325	0.6480	0.6459	0.7860
PQR (by PQA500)	0.6464	0.7300	0.7455	0.6456
DMOS (by PQA500)	0.6426	0.7296	0.7427	0.6445
Prop	0.7544	0.8072	0.8298	0.6624

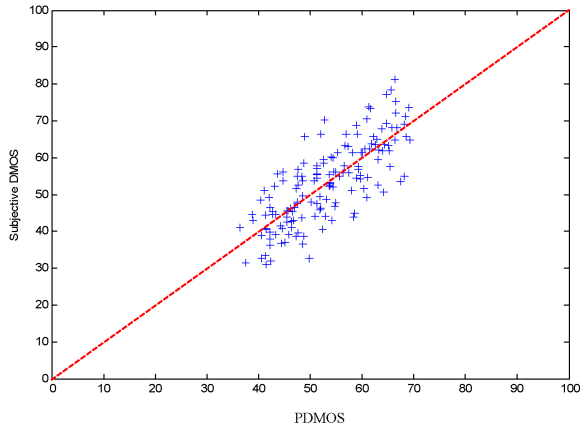


Fig. 6. Scatter plot of subjective DMOS against predicted DMOS by the proposed metric.

Table II shows the results on the LIVE Video Quality Database. Test results on the VQEG database are given in Table III. Furthermore, Table IV shows the results on each kind of distortion in LIVE database, which demonstrate that the proposed metric is rather robust to various types of video distortions.

It is impressive that the proposed metric significantly outperforms other metrics on the LIVE database according to the two indicators, and is competitive with the MOVIE index. The PSNR performs especially poorly on this database. On the VQEG database, the performance of the proposed metric is indistinguishable from other leading metrics, such as the MOVIE index, the SW-SSIM metric and the MC-SSIM metric. One important reason for the significant performance gap on the LIVE database is that it contains many spatio-temporally localized distortions that are introduced by packet loss. Compared to the blurring and blocky artifacts introduced by compression, the HVS is usually more intolerable to this type of distortion.

TABLE V
PERFORMANCE COMPARISON BETWEEN DOUBLE SALIENCY DETECTION AND SINGLE SALIENCY DETECTION

Mechanism	Spearman CC	Pearson CC
Double detection	0.779	0.778
Single detection	0.742	0.739

However, conventional distortion models, especially the pixel based models are usually incapable to account for this type of distortion. The experimental results also demonstrate that without explicitly calculating the optical flows, the extracted spatio-temporal structural descriptors could well account for these spatio-temporally localized structural distortions.

Fig. 6 shows the scatter plots of the DMOS (scaled to the full range of 1–100) against the objective prediction (after logistic regression) by the proposed metric (on the LIVE database). From the scatter plots graph we could conjecture that the proposed metric performs well from low quality cases to high quality cases.

B. Complexity Analysis

An important superiority of our algorithm is its complexity. The computation cost of the algorithm mainly concentrates on three steps: gradients calculation, 3-D structure tensor construction, and eigenvalue decomposition of these tensors.

It can easily be shown that the computational complexity of Sobel gradients computation is $O(MN_{\text{Sobel}}{}^3)$, where M is the number of pixels in one frame and N_{Sobel} is the kernel size of the Sobel operator, which is 3 in our work. The computational complexity of 3-D structure tensor construction is $O(M'N_W^2)$, where M' is the number of selected saliency pixels in one frame, and N_W is the window size for gradients integration, whose configuration will be discussed later. Finally, eigenvalue decomposition of the tensors has a computational complexity of $O(M'N_T^3)$, where N_T is the dimension of the tensor matrix, which is 3. Therefore, the total time complexity is $O(MN_{\text{Sobel}}{}^3+M'(N_W^2+N_T^3))$.

There are mainly two parameters that affect the performance and computational complexity of the algorithm: the integration windowsize N_W and the saliency threshold ε.

For the integration windowsize N_W, we tried 3×3, 5×5, 7×7, 9×9 and 11×11. The performance is almost the same at 3×3, 5×5 and 7×7 and slightly declined at 9×9 and 11×11. This is because that too large integration window would mask some small-scale distortions and tend to rate a higher similarity. Taking the computational complexity into account, we recommend to set the integration window size as 3×3. One thing that should be noted is although the integration window for the structure tensor is 3×3, however, the elements of the matrix, the gradients, are calculated by a $3\times 3\times 3$ Sobel operator. That is to say, the support region of the data analysis is actually $5\times 5\times 3$. As illustrated in Fig. 7, for the central pixel in the middle frame, we need gradients of nine pixels in the blue square to construct its structure tensor. For each pixel in the blue square (we take the up-left pixel as example), spatial and temporal gradients are calculated using the $3\times 3\times 3$ pixels in the green squares. Therefore, the descriptors describe

TABLE VI
SUMMARY OF TIME COMPLEXITY OF THE PROPOSED ALGORITHM AND OPTICAL FLOW CALCULATING ALGORITHMS

Algorithm	Proposed VQA algorithm	Optical flow algorithms		
		Phase based method	Block matching method	LK method
Time complexity	$O(MN_{\text{Sobel}}{}^3+M'(N_W{}^2+N_T{}^3))$	$O(MV_{\max}{}^3)$	$O(MV_{\max}{}^2)$	$O(M(n^2N_P+N_P^3))$

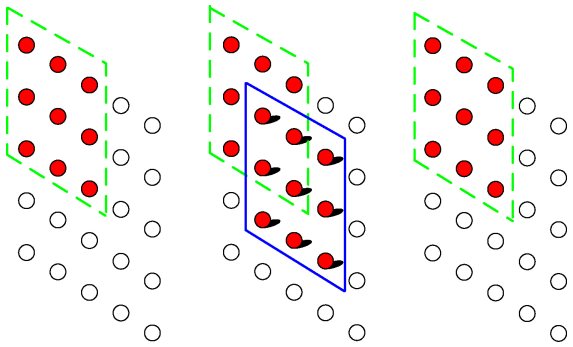


Fig. 7. Illustration for the support region of the structural descriptors. For the central pixel in the middle frame, pixels in the blue square are used to construct the structure tensor. Pixels in the green squares are used to calculate spatial and temporal gradients of the up-left pixel in the blue square. Therefore, structural descriptors of the central pixel in the middle frame summarize the energy distribution within the whole $5 \times 5 \times 3$ region in the figure.

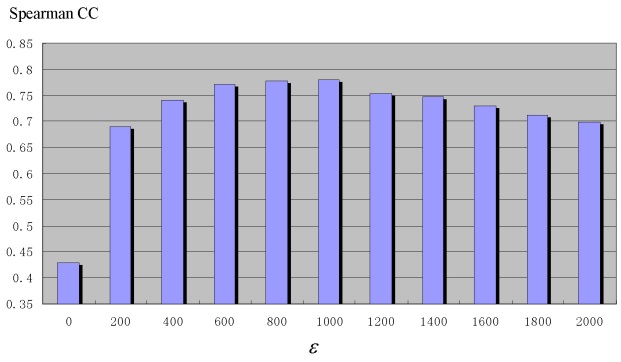


Fig. 8. Performance of the metric versus different thresholds on the LIVE database.

the energy distribution within this $5 \times 5 \times 3$ region. This size is capable to reflect the structural characteristics of natural scene primitives as well as the structural distortion cased by block based video compression and processing.

The threshold value ε for saliency detection plays an important role for both performance and computational complexity. An appropriate threshold value can not only extract the HVS-sensitive regions of the video, but also maintain the computational complexity at a proper level.

Fig. 8 illustrates the performance of the metric versus different thresholds in terms of Spearman CC coefficient (on the LIVE database). We can see the metric gives the best performance when ε is set around 1000. When $\varepsilon = 0$, all pixels are included in the evaluation, and the performance is quite poor. The efficiency of the metric will also be decreased as the threshold increases, since some important video details may be missed.

Table V shows the performance comparison between the mechanisms of double saliency detection and single saliency detection ($\varepsilon = 1000$) on LIVE database. The efficiency of the proposed method is demonstrated. With this threshold, the M' is averagely reduced to 1/10 of M in the LIVE database. As a result, computing cost is largely cut down.

As for the memory requirement, only three frames need to be kept during processing, and intermediate results will occupy only a little space.

In comparison, we will analyze the computational complexity of other optical flow based VQA algorithms. Generally, the most time-consuming part of these algorithms is the calculation of optical flow. Therefore, here we only take the complexity of optical flow calculation as a comparison. As we know, there are quite a lot of algorithms for optical flow calculation and algorithms with higher accuracy usually needs higher computing cost. The MOVIE index [8] utilized the phase based method [34] proposed by Fleet and Jepson, which is one of the most accurate methods but with the highest computing cost. Time complexity of this method is $O(MV_{\max}{}^3)$ [35], where $V_{\max}$ is the maximum motion velocity with the magnitude of dozens. The motion based SSIM [10] adopts the simplest block-matching based motion estimation method, whose complexity is $O(MV_{\max}{}^2)$. However, this method could not generate dense motion field and is unable to handle complex motion such as rotating and deforming. Another famous method is the gradient based LK method [36], which has a good tradeoff between performance and complexity. Its complexity is $O(M(n^2N_P+N_P^3))$ [37], where n is the size of a spatial template and N_P is the number of warping parameters. Usually n is set as 5, and 6-parameter affine model is used for warping. Note that 3-D structure tensor is implicitly present in this method, which also reflects the close relationship between the 3-D structure tensor and video motion characteristics. See Table VI for a summary of the algorithms' time complexity.

We compare the processing speed of the proposed algorithm with MOVIE index on the following computer: MS Windows XP professional, Inter Core2 CPU E6600 at 2.4 GHz, 3 GB of RAM. The source code is written in C, without any optimization through parallel processing or assembly language, such as MMX or SSE2 instruction. Code of MOVIE index is downloaded from [38]. The processing speed of the proposed algorithm achieves 10 frames/s on the LIVE database, while running time of the MOVIE index is about 3000 times slower.

From the above analysis, we could see that the proposed algorithm has a significant superiority in terms of computational complexity, which makes it more practical in realtime applications.

IV. Conclusion

In this paper, a novel video quality metric based on local spatio-temporal structural characteristic was proposed. Video quality was evaluated by computing the difference between two descriptors which are extracted from local spatio-temporal cubes. Experimental results on the LIVE database and VQEG FRTV Phase I database showed that the proposed metric outperforms conventional quality metrics such as PSNR, SSIM and performs competitively with MOVIE metric. Moreover, the experimental results also showed that the proposed metric was rather robust to various types of video distortions and the performance was not parameter-dependent.

References

[1] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, ITU-T Rec. J.144, Recommendations of the ITU, Telecommunication Standardization Sector, Mar. 2004.

[2] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[3] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993, pp. 207–220.

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

[5] Z. Wang, L. G. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Jan. 2004.

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

[7] M. 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.

[8] 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.

[9] A. Ninassi, O. L. Meur, P. L. 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.

[10] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1653–1658, Nov. 2010.

[11] A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 422–429, Feb. 2006.

[12] D.-O. Kim and R.-H. Park, "New image quality metric using the Harris response," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 616–619, Jun. 2009.

[13] H.-S. Han and D.-O. Kim, "Structural information-based image quality assessment using LU factorization," *IEEE Trans. Consumer Electron.*, vol. 55, no. 1, pp. 165–171, Jan. 2009.

[14] D. L. Donoho and A. G. Flesia, "Can recent innovations in harmonic analysis 'Explain' key findings in natural image statistics," *Netw. Comput. Neural Syst.*, vol. 12, no. 3, pp. 371–93, Aug. 2001.

[15] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[16] X. Li, "Edge directed statistical inference and its applications to image processing," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, 2000.

[17] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations* (Applied Mathematical Sciences, vol. 147). Berlin, Germany: Springer-Verlag, 2002.

[18] B. Jahne, H. Haussocker, and P. Geissler, *Handhook of Computer Vision and Application*. Waltham, MA: Academic Press, 1999.

[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.

[20] J. Bigun, G. H. Granlund, and J. Wiklund, "Multidimensional orientation estimation with applications to texture analysis and optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 775–790, Aug. 1991.

[21] J. Weickert, *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner, 1998.

[22] B. Jähne, "Spatio-temporal image processing," in *Lecture Notes in Computer Science*, vol. 751. Berlin, Germany: Springer, 1993.

[23] T. Brox, J. Weickert, B. Burgeth, and P. Mrázek, "Nonlinear structure tensors," *Image Vision Comput.*, vol. 24, no. 1, pp. 41–55, 2006.

[24] J. Zhang, J. Gao, and W. Liu, "Image sequence segmentation using 3-D structure tensor and curve evolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 5, pp. 629–641, May 2001.

[25] I. T. Jolliffe, *Principle Component Analysis*. New York: Springer-Verlag, 1986.

[26] L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, and W. Gao, "No-reference perceptual image quality metric using gradient profiles for JPEG2000," *Sig. Proc.: Image Comm.*, vol. 25, no. 7, pp. 502–516, 2010.

[27] Z. Lu, W. Lin, X. Yang, E.-P. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[28] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 2560–2563.

[29] L. Itti and C. Koch "Computational modeling of visual attention," *Nat. Rev. Neurosci.*, vol. 3, no. 2, pp. 194–203, Mar. 2001.

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Jason Hopkins University Press, 1983.

[31] 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.

[32] VQEG: The Video Quality Experts Group. (2000). [Online]. Available: http://www.its.bldrdoc.gov/vqeg

[33] A. M. Rohaly, J. Libert, P. Corriveau, and A. Webster. (2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment* [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtvphaseI

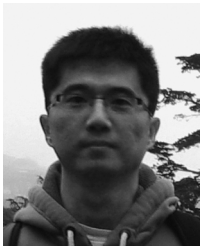
[34] D. J. Fleet and A. L. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77–104, 1990.

[35] L. Hongche, T.-H. Hong, M. Herman, and R. Chellappa, "Accuracy vs. efficiency trade-offs in optical flow algorithms," in *Proc. ECCV*, vol. 2. 1996, pp. 174–183.

[36] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[37] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

[38] Software Release. (2010). *Motion-Based Video Integrity Evaluation (MOVIE) Index* [Online]. Available: http://live.ece.utexas.edu/research/quality/movie.html



Yue Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006. He is currently pursuing the Ph.D. degree with the Graduate University of the Chinese Academy of Sciences, Beijing.

His current research interests include image and video coding, processing, and quality assessment.

Tingting Jiang received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2001, and the Ph.D. degree in computer science from Duke University, Durham, NC, in 2007.

She is currently an Assistant Professor of computer science with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. Her current research interests include computer vision, image, and video quality assessment.

Siwei Ma (S'03) received the B.S. degree in computer science from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

From 2005 to 2007, he was a Post-Doctorate with the University of Southern California, Los Angeles. Then, he joined the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, where he is currently an Associate Professor. He has published over 30 technical articles in refereed journals and proceedings in the areas of image and video coding, video processing, video streaming, and transmission. His current research interests include image and video coding, video streaming, and transmission.

Wen Gao (M'92–SM'05–F'09) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1985, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He was a Professor in computer science with the Harbin Institute of Technology from 1991 to 1995, and a Professor in computer science with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, from 1996 to 2005. He is currently a Professor with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing. He has been leading research efforts to develop systems and technologies for video coding, face recognition, sign language recognition and synthesis, and multimedia retrieval. He has published four books and over 500 technical articles in refereed journals and proceedings in the areas of signal processing, image and video communication, computer vision, multimodal interface, pattern recognition, and bioinformatics. His current research interests include signal processing, image and video communication, computer vision, and artificial intelligence.

Dr. Gao has received many awards, including five national awards for research achievements and activities. He did many services for the academic society, such as being the General Co-Chair of the IEEE International Conference on Multimedia and Expo in 2007 and the Head of Chinese delegation to the Moving Picture Expert Group of International Standard Organization since 1997. He is also the Chairman of the working group responsible for setting a National Audio Video Coding Standard for China.