To appear in Journal of the Optical Society of America A

Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception

Zhou Wang and Qiang Li

Dept. of Electrical & Computer Engineering, University of Waterloo, ON N2L3G1, Canada
Dept. of Electrical Engineering, The University of Texas at Arlington, TX 76019, USA

zhouwang@ieee.org; qiangli@uta.edu

Motion is one of the most important types of information contained in natural video, but direct use of motion information in the design of video quality assessment algorithms has not been deeply investigated. Here we propose to incorporate a recent model of human visual speed perception [Stocker & Simoncelli, *Nature Neuroscience* **9**, 578-585 (2006)] and model visual perception in an information communication framework. This allows us to estimate both the motion information content and the perceptual uncertainty in video signals. Improved video quality assessment algorithms are obtained by incorporating the model as spatiotemporal weighting factors, where the weight increases with the information content and decreases with the perceptual uncertainty. Consistent improvement over existing video quality assessment algorithms is observed in our validation with the video quality experts group Phase I test data set.

OCIS codes: 110.3000, 330.5510, 330.4060, 100.2960.

1. Introduction

The capability of representing motion is probably the most critical characteristic that distinguishes a natural video sequence from a stack of independent still image frames. If we believe that the central goal of vision is to extract useful information from the visual scene, then the perception of motion information would play an important role in the perception of natural video. Since the main purpose of objective video quality assessment (VQA) is to predict human behavior in the evaluation of video quality, it would be essential for a successful VQA system to effectively take into account motion information.

Nevertheless, in the literature of VQA, motion information has typically been employed *indirectly*. The most frequently used method is temporal filtering [1, 2], where linear filters or filter banks are applied along the temporal direction (or along the spatial and temporal directions simultaneously), and the filtered signals are normalized to reflect the effect of the temporal contrast sensitivity function [3] (the variation of human visual sensitivity as a function of temporal frequency). Advanced models may also include the temporal masking effects (the reduction of visibility of one image component due to the existence of its neighboring components) [2] or statistics of the temporal filter coefficients [4]. Since motion in the visual scene may cause variations in signal intensity along the temporal direction, temporal filtering can, to some extent, capture motion. However, representing motion using temporal filtering responses is indirect, inaccurate, and in some sense problematic. First, motion may not be the sole reason for temporal signal intensity variations. The change of lighting conditions is an obvious counterexample. Therefore, the temporal filter coefficients are indeed a mixture effect of motion together with many other reasons. Second, the speed of motion cannot be directly related to the strength of temporal filter responses. For example, two objects with the same speed of motion but different texture and contrast would result in different speeds of temporal intensity variation, and thus different temporal filter responses. Third, many visual experiments that measure temporal visual sensitivities were done with flickering patterns [1], which do not reflect any physical motion of the objects. Moreover, since the motion and speed information are not represented explicitly, a lot of knowledge about motion perception cannot be directly used within such a temporal filtering framework.

Only a relatively small number of existing VQA algorithms detect motion explicitly and use motion information directly. Wang *et al.* proposed a heuristic weighting model [5], which was combined with the structural similarity (SSIM) [6] based quality assessment method to take into account the fact that the accuracy of visual perception is significantly reduced when the speed of motion is extremely large. A set of heuristic fuzzy rules was proposed by Lu *et al.* [7] that use both absolute and relative motion information to account for visual attention and motion suppression. It was shown that these rules are effective in improving VQA performance of the standard mean squared error (MSE)/peak signal-to-noise ratio (PSNR) measures as well as the SSIM [6] approach. In two recent papers by Seshadrinathan and Bovik, local motion information obtained from optical flow computation is employed to adaptively guide the orientation of a set of three-dimensional Gabor filters [8,9]. The adapted Gabor filter responses are then incorporated into the SSIM [6, 10] and the visual information fidelity (VIF) [11] measures for the purpose of VQA.

In this paper, we propose what we believe to be a new method that directly incorporates motion information by modeling the visual perception process in an information communication framework. Our approach is largely inspired by the recent psychophysical study by

Stocker and Simoncelli on human visual speed perception [12]. Based on a Bayesian optimal observer hypothesis, Stocker and Simoncelli measured the prior probability distribution and the likelihood function of speed perception simultaneously from a set of carefully designed psychovisual experiments. These measurements are consistent across human subjects and can be modeled using simple parametric functions. These results are substantially different from previous statistical models of visual speed perception [13–15], where the prior distributions are assumed rather than measured. Our approach has greatly benefited from these results, because the statistical models derived from them provide the essential ingredients in the computation of both the motion information content and the perceptual uncertainty (details will be given in Section 2). Our method is based on the following assumptions and observations.

First, we believe that the human visual system (HVS) is an efficient encoder or information extractor (subject to certain physical constraints such as power consumption), as widely hypothesized in computational vision science [16]. To achieve such efficiency, it is natural to assume that the areas in the visual scene that contain more information should be more likely to attract visual attention and fixations [17, 18]. Such *information content* can be quantified using statistical information theory, provided that a statistical model about the information source is available. In fact, the information content-based method has already shown to be useful in still image quality assessment (IQA) [19].

Second, as in a number of previous papers [4, 11, 19], we model visual perception as an information communication process, where the information source (the video signal) passes through an error-prone communication channel (the HVS). The key difference from the previous IQA/VQA models is that the noise level in the communication channel is not fixed here. This is motivated by the empirical observation that the HVS does not perceive all the information content with the same degree of certainty. For example, when the background motion in a video sequence is very large (or the head/camera motion is very large), the HVS cannot identify the objects presented in the video with the same accuracy as in static background images, i.e., the video signal is perceived with higher uncertainty. Again, such *perceptual uncertainty* can be quantified based on information theory, by relating the stochastic channel distortion model with the speed of motion. In particular, the psychophysical study by Stocker and Simoncelli [12] suggests that the internal noise of human visual speed perception increases with the true stimulus speed and decreases with the stimulus contrast.

In Section 2, we describe our method to compute locally (in both space and time) the information content and the perceptual uncertainty based on the motion information estimated from the video sequence. We then combine them to generate a three-dimensional perceptual weight function. The function can be incorporated as weighting factors into any local VQA

algorithm that produces a quality/distortion map over space and time. More detailed implementation is presented in Section 3. In Section 4, we validate our model by combining it with MSE/PSNR and SSIM [6] based quality assessment methods. Consistent improvement is achieved with the video quality experts group (VQEG) Phase I test database [21].

2. Method

The motion information in a video sequence can be represented as a three-dimensional field of motion vectors, where each spatial location (x, y) and time instance t is associated with a motion vector $\vec{v}(x, y, t) = [v_x(x, y, t)\ v_y(x, y, t)]^T$. For notational convenience, in the rest of the paper, we often drop the space and time indices and write a motion vector as $\vec{v}$. For a given video sequence, we consider three types of motion fields − absolute motion, background motion, and relative motion. An illustration is given in Fig. 1, where the absolute motion $\vec{v}_a$ is estimated as the absolute pixel movement at each spatial location between two adjacent video frames. By contrast, the background motion $\vec{v}_g$ is approximately global, which is often caused by the movement of the image acquisition system. We also define a relative motion $\vec{v}_r$ at each spatial location as the vector difference between the absolute and the global motion, i.e.,

$$\vec{v}_r = \vec{v}_a - \vec{v}_g\,. \tag{1}$$

The speed of motion can be computed as the length of the motion vector, which, for convenience, we denote as $v = \|\vec{v}\|_2$. Thus, v_g, v_a and v_r represent the speed of the background motion, the absolute motion, and the relative motion, respectively.

A recent approach in understanding human visual speed perception is to use a Bayesian optimal observer model, in which the visual system judges the speed of motion by "optimally" combining some prior knowledge of the visual world together with the current noisy measurements [12–14]. It has been shown that this approach can successfully explain a number of psychovisual phenomena where the visual system tends to give biased judgments on the speed of retinal motion [12–14]. Figure 2 describes this approach in an information communication framework, where the stimulus speed information v passes through a noisy front-HVS channel. This results in the internal noisy measurement m, which is associated with a statistical noise model, or a likelihood function. The visual system gives an estimate of the stimulus speed $\hat{v}$ not only from m, but also based on some prior information about the probability distribution of the stimulus speed. It is assumed that the prior distribution has been established beforehand in the brain by sufficient statistics about the natural visual environment. Figure 2 also shows the prior distribution and the noise likelihood function measured by Stocker and Simoncelli [12]. In this section, we will describe how these measurements help us develop models to compute both the information content and the perceptual uncertainty of speed perception and how to combine them for VQA.

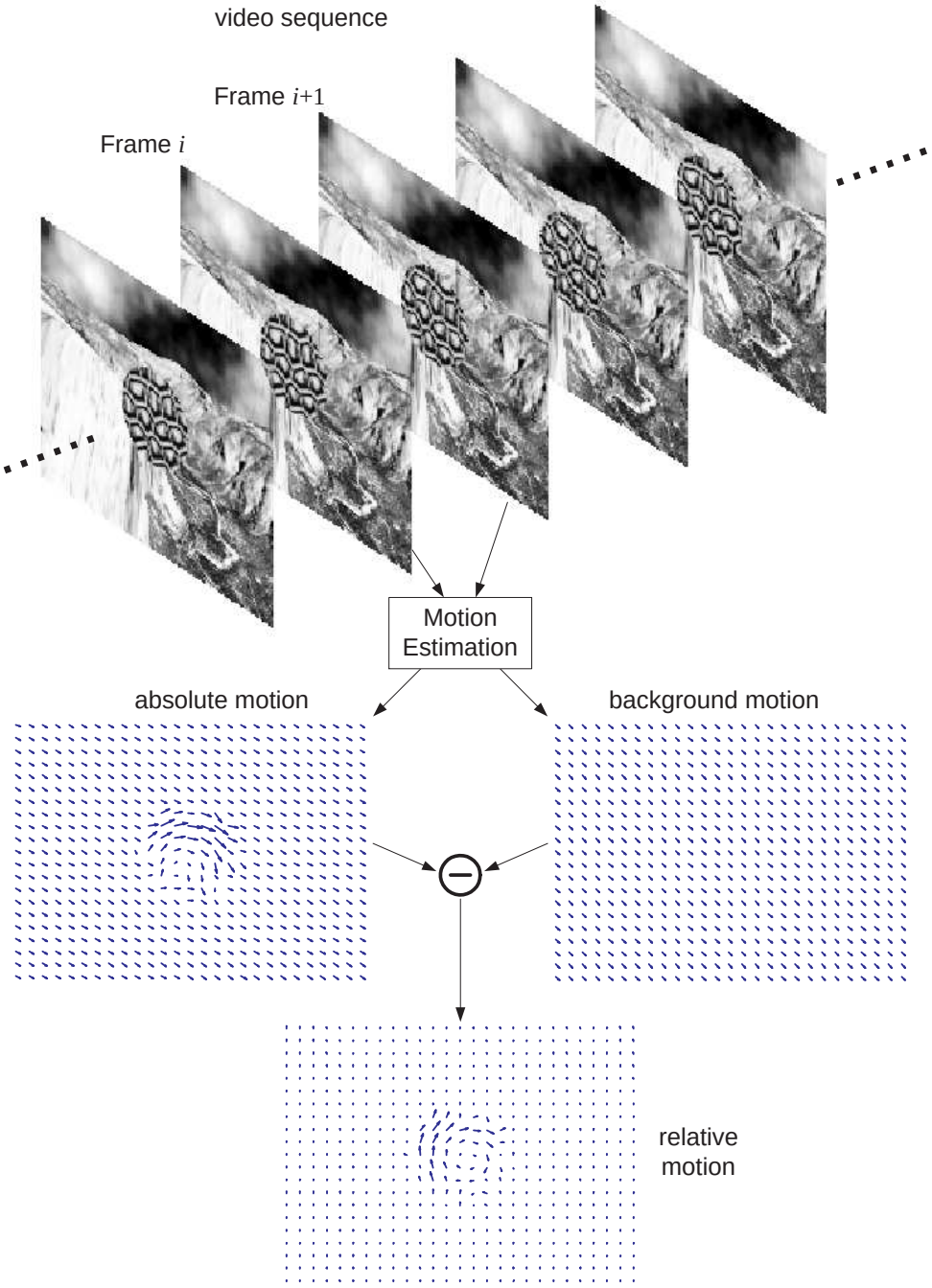

Fig. 1. Illustration of absolute motion, background motion and relative motion estimated from two consecutive frames of a video sequence.

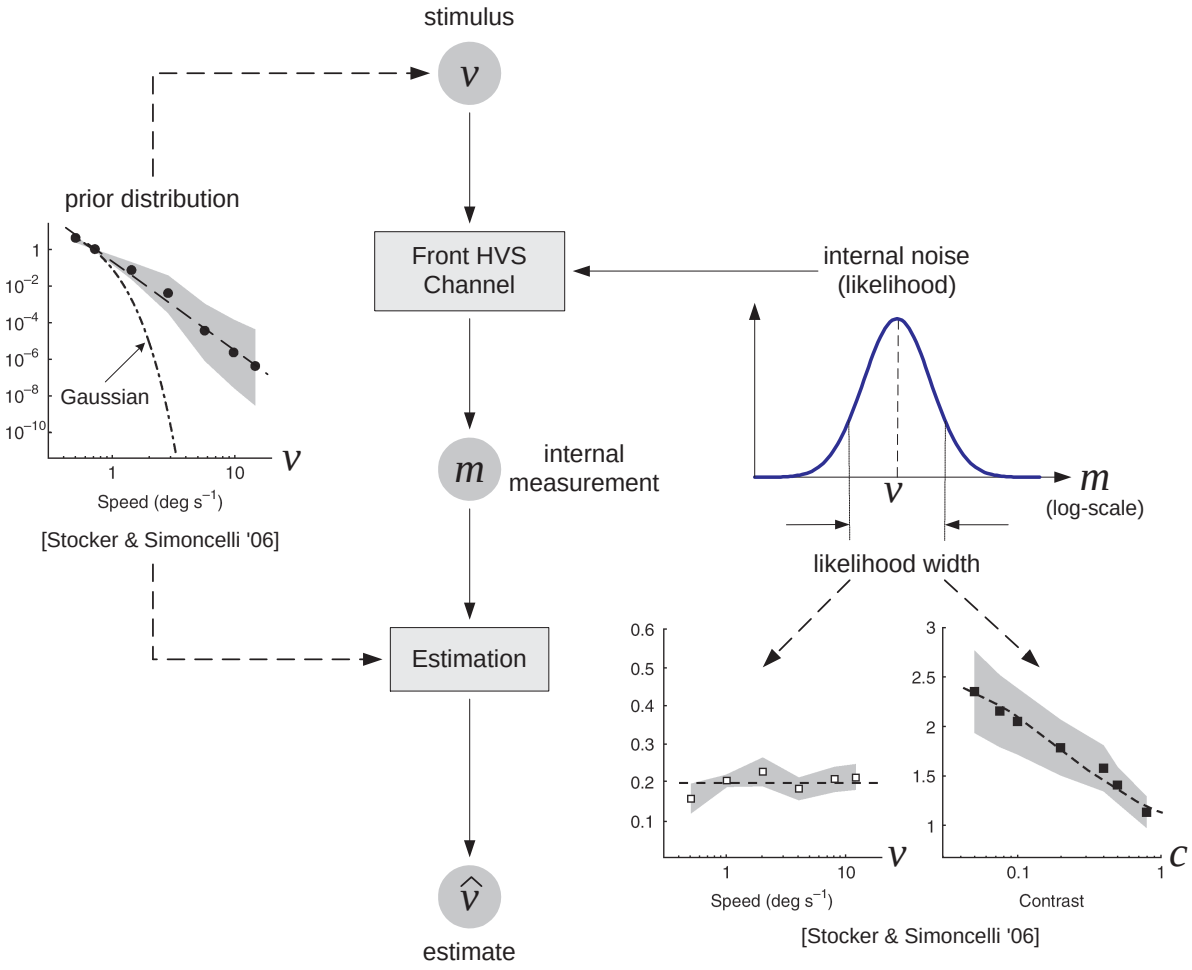

Fig. 2. Bayesian visual speed perception in an information communication framework. v: stimulus speed; m: noisy measurement; $\hat{v}$: estimated speed; c: stimulus contrast. Adapted from [Stocker & Simoncelli '06] [12].

2.A. *Information Content*

It is believed that object motion is associated with visual attention and can be used for predicting visual fixations [20]. This is intuitively sensible because statistically, most of the objects in the visual world are static (or close to static) relative to the background. As a result, an object with significant motion with respect to the background would be a strong *surprisal* to the visual system. If the HVS is an efficient information extractor, as discussed in Section 1, then it should pay more attention to such a surprising event. This intuitive idea may be converted into a quantitative measure of motion information content (or how surprising the event is), provided that the prior probability distribution about the speed of motion is known. Early work on Bayesian speed perception has assumed Gaussian distribution for the speed prior [13], but the recent result by Stocker and Simoncelli [12] suggests that the distribution has a much longer tail than Gaussian, as shown in Fig. 2. Indeed, it can be well fitted with a straight line in the log-log domain (see Fig. 2). This leads us to assume a power-law function for the prior distribution of relative motion:

$$p(v_r) = \frac{\tau}{v_r^{\alpha}}, \tag{2}$$

where τ and α are two positive constants. Since the power-law function does not sum to a finite number, this is not a strictly valid probability density function and can only be used when v_r is away from 0. For any observed motion v_r, we can then estimate the information content associated with it by computing its self-information or surprisal as

$$I = -\log p(v_r) = \alpha \log v_r + \beta, \tag{3}$$

where $\beta = -\log \tau$ is a constant. Eq. (3) suggests that the motion information content increases with the speed of relative motion, which is consistent with our intuition discussed earlier.

2.B. *Perception Uncertainty*

If we model visual perception as an information communication process, then the amount of information that can be received (perceived) at the receiver end will largely depend on the noise in the distortion channel (the HVS). In other words, the internal noise in the HVS, or the likelihood function of the noisy measurement, determines the perceptual uncertainty. It was found that for a given stimulus speed, a log-normal distribution can provide a good description of the likelihood function [12]:

$$p(m|v_s) = \frac{1}{\sqrt{2\pi}\sigma m} \exp\left[\frac{-(\log m - \log v_s)^2}{2\sigma^2}\right], \tag{4}$$

where v_s and m are the speed of the true stimulus motion and the measurement, respectively. Furthermore, the experimental results by Stocker and Simoncelli [12] suggest that in the

logarithmic speed domain, the width parameter σ in the log-normal distribution is roughly constant for any stimulus speed v_s and inversely dependent on the stimulus contrast c, as illustrated in Fig. 2. Note that the width here is represented in the log-domain, and thus it indeed scales linearly with v_s in the linear speed domain. Mathematically, we model it as

$$\sigma = \frac{\lambda}{c^{\gamma}}\,, \tag{5}$$

where λ and γ are both positive constants.

For a given video sequence, we assume that the underlying stimulus speed v_s is the speed of the background motion v_g. This assumption is naturally connected to our intuitive idea described in Section 1 that when the background motion in a video sequence is very large (most likely caused by large head/camera motion), the HVS cannot identify the objects presented in the video with the same accuracy as in static background. A natural way to quantify the level of the internal noise, or the perceptual uncertainty, is the entropy of the likelihood function, which can be computed as

$$\begin{aligned} U &= -\int_{-\infty}^{\infty} p(m|v_g)\log p(m|v_g)dm \\ &= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2) + \log v_g \\ &= \log v_g - \gamma\log c + \delta\,, \end{aligned} \tag{6}$$

where $\delta = \frac{1}{2} + \frac{1}{2}\log(2\pi) + \log\lambda$ is a constant. Again, this perceptual uncertainty measurement is consistent with our intuition. On the one hand, it increases with the background motion of the video frame, suggesting that when the background motion is very large, the HVS cannot extract the structural information about the objects presented in the video with the same accuracy as in static images. On the other hand, it decreases with the stimulus contrast, implying that higher contrast objects are perceived with lower uncertainty.

2.C. Video Quality Assessment Based on Motion Perception

We compute the motion information content and the perceptual uncertainty at every spatial location and time instance (x, y, t) in the video sequence. Based on the efficient coding hypothesis about the HVS, the importance of a visual event should increase with the information content, and decrease with the perceptual uncertainty. Therefore, we define the following spatiotemporal importance weight function at every (x, y, t)

$$w = I - U = (\alpha\log v_r + \beta) - (\log v_g - \gamma\log c + \delta)\,. \tag{7}$$

The calculation of the information content, the perceptual uncertainty and the importance weighting function is demonstrated in Fig. 3, where two consecutive video frames are extracted from the "Mobile Calendar" sequence, and the motion field as well as the maps for

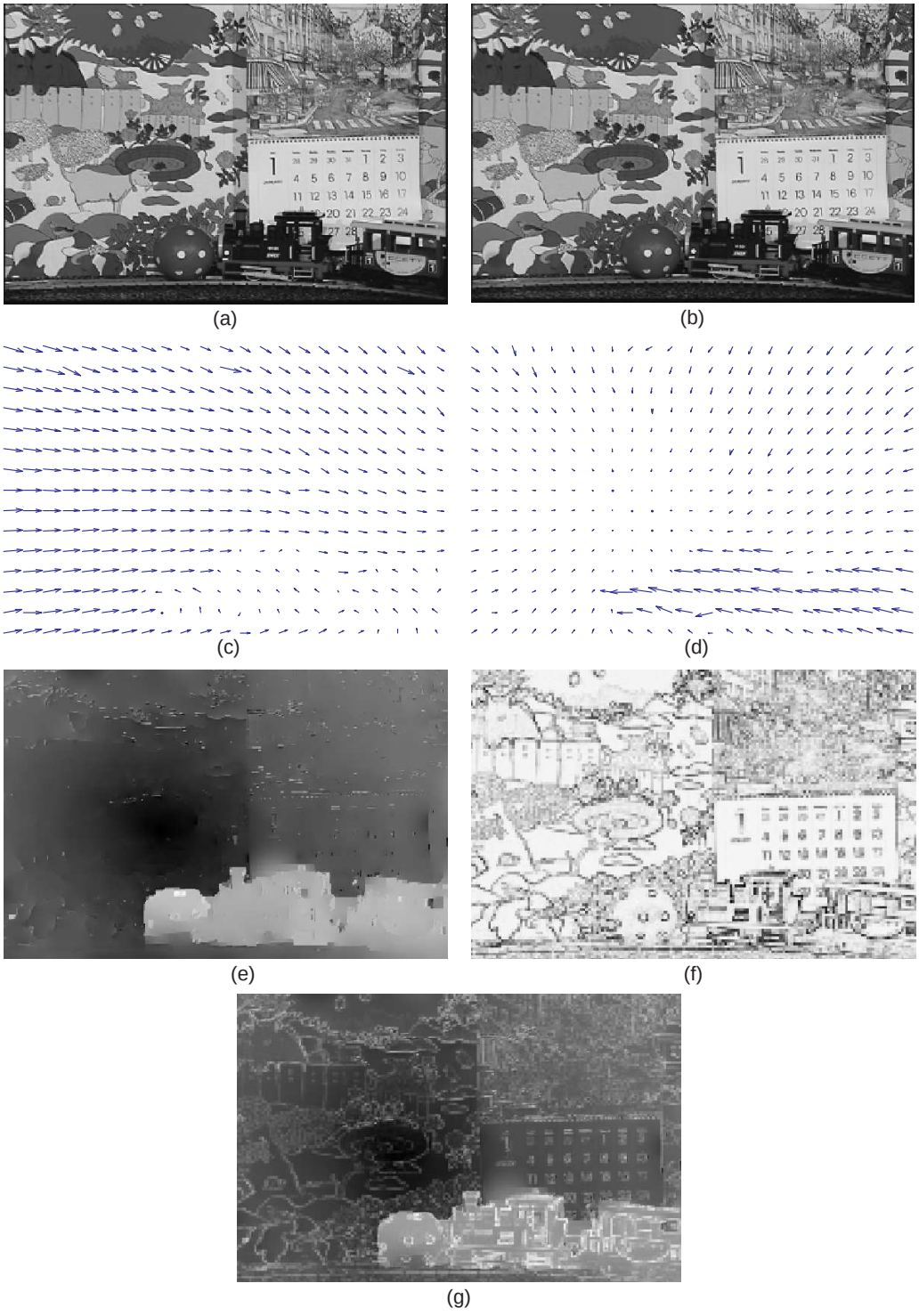

Fig. 3. (a),(b) Two consecutive frames extracted from the "Mobile Calendar" sequence; (c) Estimated absolute motion field; (d) Estimated relative motion field; (e) Estimated local information content map; (f) Estimated local perceptual uncertainty map; (g) Estimated local weighting factor map.

I, U and w are computed (more details of the computation are given in Section 3). It is observed from the video sequence that the toy train moves from the right to the left with respect to the moving background. Note that although the absolute motion of the train is almost static, its relative motion is significant. Thus, based on our model, the region associated with the train is given larger weights relative to the background.

The importance weight function alone cannot serve as a VQA algorithm. However, it can be incorporated into a local image quality/distortion measure as a weighting function. The local image quality/distortion measure must provide a three-dimensional quality/distortion map of the video sequence being evaluated. Let $q(x, y, t)$ be the quality/distortion map given by the local quality/distortion metric, the final VQA score is computed as

$$Q = \frac{\sum_t \sum_x \sum_y w(x, y, t)\, q(x, y, t)}{\sum_t \sum_x \sum_y w(x, y, t)}\,. \tag{8}$$

3. Implementation Issues

To build a real VQA system based on the proposed approach, several implementation issues need to be resolved. First, we need to estimate the motion vector field. Rather than using block matching-based motion estimation as in previous work [5], here we choose to use a optical flow method for motion estimation, which avoids the computationally intensive block search procedures and provides a smoother motion vector field. In particular, we compute the absolute motion field using Black and Anandan's multi-layer optical flow estimation algorithm [22] with a five-level pyramid decomposition. The background motion is obtained by a maximum likelihood estimation to identify the peak of the histogram generated by histogramming motion vectors on the two-dimensional grid [23]. The relative motion vector $\vec{v}_r$ is then computed using Eq. (1).

Second, the local contrast needs to be computed at each spatial location and time instance. Although contrast is an extensively used term throughout the field of visual psychophysics and physiology, mathematical definition of local contrast for complex natural images is a nontrivial issue [24]. Here we compute the local contrast as the ratio between the local standard deviation normalized by the local mean, i.e., for a given local image patch p (8×8 blocks in our implementation), we define

$$c' = \frac{\sigma_p}{\mu_p + \mu_0}\,, \tag{9}$$

where σ_p and μ_p are the standard deviation and the mean computed within the local patch, respectively, and μ_0 is a small constant to avoid instability near 0. This definition of contrast guarantees that linearly scaling all the pixel intensities around the mean leads to a linear scaling in the contrast calculation. Compared to the choice of using the difference between the maximal and minimal pixel intensities, this contrast definition also avoids the instability

that one extreme pixel (e.g., a positive or negative impulse) could drastically change the contrast evaluation of an image patch. In addition, as in previous models [25, 26], to take into account the contrast response saturation effect at small and large contrast values, we pass the contrast computation through a pointwise nonlinear function given by

$$c = 1 - e^{-(c'/\theta)^\rho}, \tag{10}$$

where ρ and θ are two constants that control the slope and the position of the function, respectively.

The third practical issue in the implementation of the algorithm is that the background motion v_g, the relative motion v_r, and the local contrast c may be close to zero. This could result in unstable evaluation of the weight function. To avoid this, and to take into account the Weber-Fechner law, we take a similar approach as in the Stocker and Simoncelli paper [12]. That is, instead of computing $\log v_r$, $\log v_b$, and $\log c$, we replace them with $\log(1 + v_r/v_0)$, $\log(1 + v_b/v_0)$, and $\log(1 + c/c_0)$, respectively, where v_0 and c_0 are both small positive constants. Furthermore, to avoid the situation that the weight might go negative, we threshold it at 0. Therefore, the final importance weight function we are computing is given by

$$w = \max\left\{0, \left[\alpha \log\left(1 + \frac{v_r}{v_0}\right) + \beta\right] - \left[\log\left(1 + \frac{v_g}{v_0}\right) - \gamma \log\left(1 + \frac{c}{c_0}\right) + \delta\right]\right\}. \tag{11}$$

Since the motion vectors are in the unit of pixels/frame, the parameter v_0 also needs to be in the same unit. In our implementation, we assume a 32 pixels/degree of viewing distance, and as in the Stocker and Simoncelli paper, we fix $v_0 = 0.3$ degree/sec. We can then convert v_0 based on the frame rate of the video sequence. For example, if the frame rate is 30 frames/sec, then $v_0 = 0.3 \times 32/30 = 0.32$ pixels/frame. If the frame rate is 25 frames/sec, then $v_0 = 0.3 \times 32/25 = 0.384$ pixels/frame. The other parameters are handpicked and we find that the following parameters give reasonable results and use them in all the experiments reported later in this paper: $\alpha = 0.2$, $\beta = 0.09$, $\gamma = 2.5$, $\delta = 2.25$, $\mu_0 = 6$, $\theta = 0.05$, $\rho = 2$, and $c_0 = 0.07$. In our experiments, we found that generally the overall performance of the algorithm is not very sensitive to small variations on these parameters. However, how to choose these parameters in a systematic way and how to quantify their sensitivities are still under investigation.

4. Validation

To validate the proposed model with real VQA algorithms, we incorporate the proposed weighting method with two types of image distortion/quality maps. The first is the squared error map defined by

$$q(x, y, t) = |I_r(x, y, t) - I_d(x, y, t)|^2, \tag{12}$$

where $I_r(x, y, t)$ and $I_d(x, y, t)$ are the pixel intensity values at spatial location (x, y) and time t in the original video sequence (as a perfect-quality reference) and the distorted video sequence (quality to be evaluated), respectively. The standard MSE measure is a simple average of such a distortion map over space and time. The standard PSNR measure (which is widely used in the image processing literature) is defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{L^2}{\text{MSE}} \right) , \tag{13}$$

where L is a constant, representing the dynamic range of image pixel intensities (e.g., for 8 bits/pixel gray-scale image, $L = 2^8 - 1 = 255$). Notice that the PSNR values do not provide any new information other than a nonlinear monotonic scaling of the MSE values. With the proposed weighting approach being taken into account, a weighted MSE measure can be computed using Eq. (8). This can then be further converted to a weighted PSNR measure using the same approach as Eq. (13).

The second type of image quality map is created using the SSIM approach [5,6]. The local SSIM value is computed using two image patches extracted from the same spatial location from the reference and the distorted images, respectively. The SSIM value is defined as [6]

$$q(x, y, t) = \frac{(2\mu_r\mu_d + C_1)(2\sigma_{rd} + C_2)}{(\mu_r^2 + \mu_d^2 + C_1)(\sigma_r^2 + \sigma_d^2 + C_2)} , \tag{14}$$

where μ_r, μ_d and σ_r, σ_d are the mean and standard deviation values of the reference and the distorted image patches, respectively. σ_{rd} is the cross correlation between the mean-removed image patches, and C_1 and C_2 are two constants. More detailed explanations and discussions about SSIM can be found in the referred papers [5, 6]. Again, the standard SSIM measure is a simple average of the SSIM map over all space and time, and a weighted SSIM measure can be computed by incorporating Eq. (14) into Eq. (8).

Different image distortion/quality maps can provide a substantially different prediction of local image quality. An example is shown in Fig. 4, where an original image is compressed with JPEG, and the absolute difference map (which is the basis for MSE/PSNR measure) and the SSIM map are computed. Both maps use brighter pixels to indicate better quality. Careful inspection of the distorted image together with the quality maps suggests that absolute error is not a good indicator of local image quality when compared with the SSIM index. For example, only the SSIM map clearly points out the blocking artifacts in the sky.

The proposed method is tested using the VQEG Phase I database [21], which, to the best of our knowledge, is the only publicly available database that contains a relatively large number of subject-rated video sequences. The database contains 20 standard definition television reference video sequences, which can be further divided into two sets of video sequences that are 50Hz (25 frames/s) and 60Hz (30 frames/s), respectively. Each reference

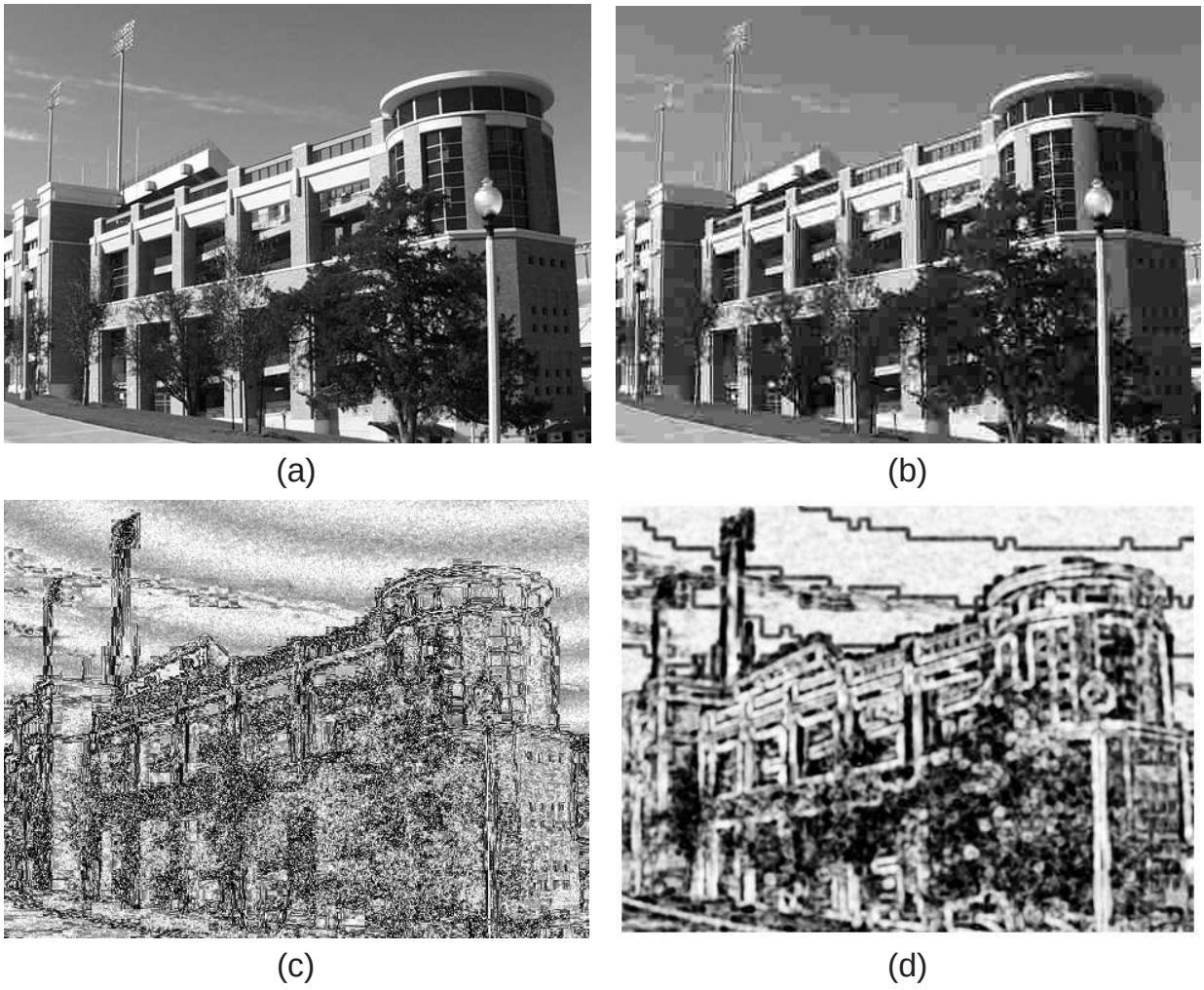

Fig. 4. Illustration of quality maps. (a) Original image; (b) distorted image (by JPEG compression); (c) absolute error map – brighter indicates better quality (smaller absolute difference); (d) SSIM index map – brighter indicates better quality (larger SSIM value). The SSIM index appears to be a better indicator of local image quality.

Table 1. ROCC Results of VQA Algorithms on VQEG Phase I Database. PSNR(sw): PSNR with spatial information content-based weighting [19]. PSNR(w): PSNR with proposed weighting; SSIM(sw): SSIM with spatial information content-based weighting [19]. SSIM(w): SSIM with proposed weighting.

Data set	MSE/PSNR	PSNR(sw) [19]	PSNR(w)	SSIM	SSIM(sw) [19]	SSIM(w)
50hz	0.8152	0.8211	0.8278	0.8301	0.8544	0.8948
60hz	0.7112	0.7120	0.7303	0.7680	0.7692	0.7985
All	0.7818	0.7887	0.8048	0.8127	0.8287	0.8621

video sequence has 16 distorted versions with a variety of distortion types [21]. This results in a total of 320 distorted video sequences. The subject score for each sequence is given by the mean opinion score (MOS) from the ratings given by multiple human subjects. The difference of MOS (DMOS) score is then calculated for each distorted video sequence by subtracting its MOS by the MOS of its corresponding reference video sequence.

We use the Spearman rank order correlation coefficient (ROCC) between the subjective and objective scores to evaluate the performance of the VQA algorithms:

$$r = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{K(K^2 - 1)}, \tag{15}$$

where K is the number of video sequences in the data set, and d_i is the difference between the i-th video sequence's ranks in subjective and objective evaluations. ROCC is one of the metrics adopted by VQEG for the evaluation of video quality measures [21]. Its advantage is in its robustness because it is independent of any fitting function that attempts to find a nonlinear mapping between the objective and the subjective scores. Table 1 shows the ROCC test results of three data sets − the 50 Hz data set, the 60 Hz data set, and all data combined. The results suggest that the proposed weighting method is quite effective. It gives clear and consistent improvement to all test data sets with two completely different types of image distortion/quality maps. Similar results are also obtained with all the other VQEG test metrics [21]. To compare the proposed approach with other visual attention based quality assessment models, we have also included the spatial information content-based attention and weighting model by Wang and Shang [19] in Table 1. It appears that this visual attention model is also helpful in improving the performance of both PSNR and SSIM (consistent improvements are observed for all data sets), but not as effective as the proposed model.

Figures 5(a), 5(b), 5(c), and 5(d) show the scatter plots of the subjective/objective com-

parisons on all VQEG test video sequences for PSNR, PSNR with proposed weighting, SSIM, and SSIM with proposed weighting, respectively. These scatter plots confirm the ROCC results shown in Table 1. It can be seen that applying the proposed weighting model has made visible impact on the tightness of the clusters of sample points (each associated with a test video sequence), which reflects the consistency between subjective and objective evaluations.

5. Discussion

The design of VQA algorithms is an important engineering problem that has a wide range of real-world applications. It is also a highly challenging problem because the ultimate purpose of VQA is to emulate the performance of the biological visual system, which is extremely complicated. In the development of VQA algorithms, it is desirable to maintain a good balance between accuracy and complexity. If there were an objective system that could simulate all related aspects of the HVS, including its built-in knowledge about the visual environment, then it should provide precise prediction of perceived video quality. However, such systems are likely to require complex implementations and intensive computations, making them cumbersome in practical applications. Therefore, a great deal of effort should be made to simplify the models without significantly losing their accuracy. We strongly believe that using high-level hypotheses about the overall behaviors of the visual system is an effective and efficient way to achieve this goal. Specifically, it seems promising to model the visual perception process in an information communication framework, and the results of a few recent IQA/VQA papers [4,11,19] as well as this paper supply initial support of this approach. Another simplification we are making in this paper is that we assume the local information content and perceptual uncertainty in a video signal are proportional to the local information content and perceptual uncertainty of speed perception. This also seems to be a useful approach to capture the important aspects of video perception while maintaining the proposed algorithm as a computationally tractable engineering solution.

One interesting observation in Table 1 is that the proposed weighting method results in larger improvement for the SSIM-based method as compared to the MSE/PSNR-based method. This is somewhat counterintuitive as the basic SSIM metric by itself has already included some extent of human visual characteristics. This is a complicated issue that is worth further investigation. One explanation might be that the absolute error map (the first step in computing MSE/PSNR) gives a very poor indication of local image quality, as exemplified by Fig. 4. As a result, even when an accurate local weighting function is applied to the map, it may not help much. Instead, sometimes it might unluckily assign large weights to the regions that the absolute error map is giving very wrong estimates about local image quality, leading to even worse results.

The proposed algorithm may be improved and/or extended in many ways. First, there

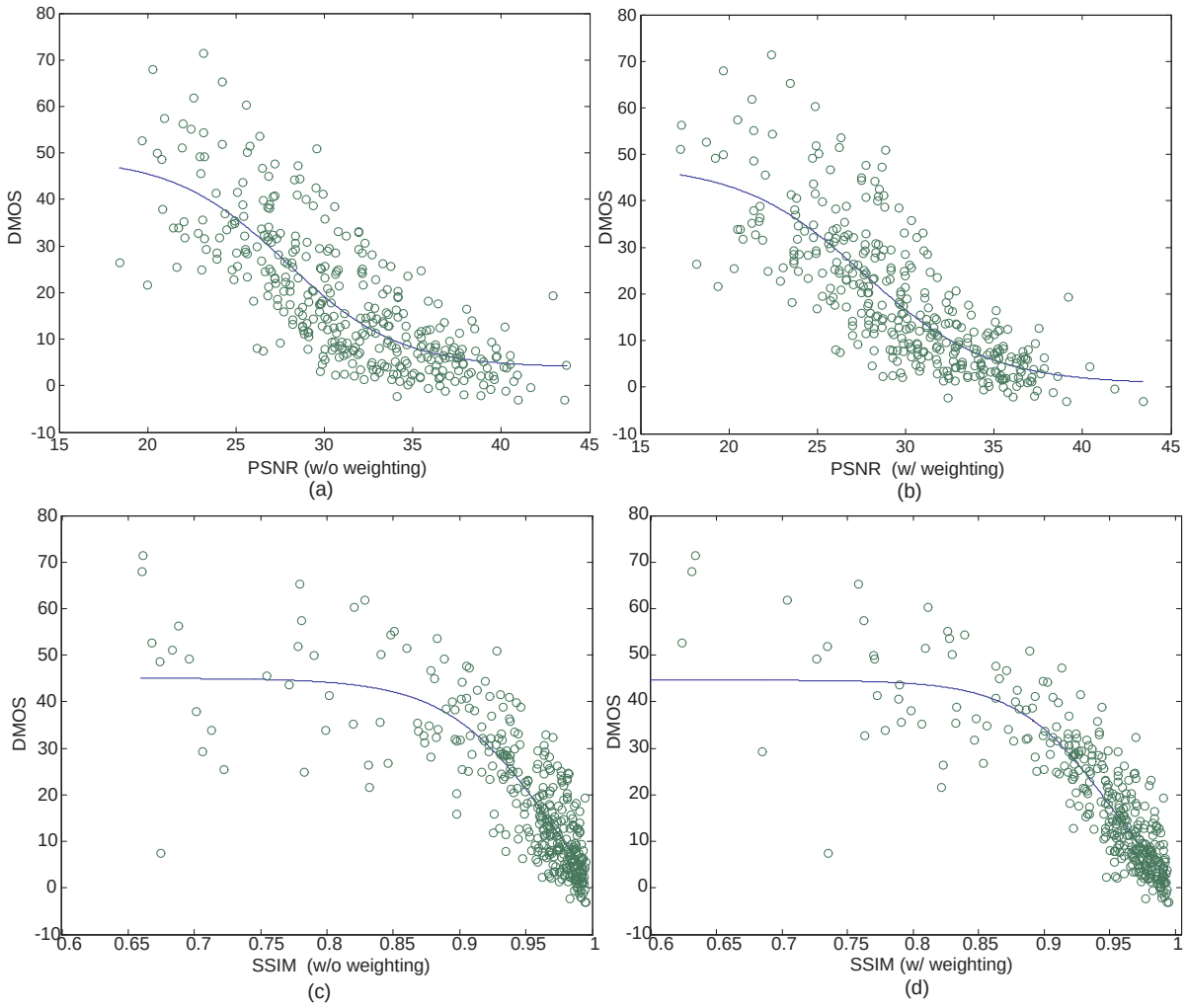

Fig. 5. Scatter plots of subjective/objective scores on VQEG Phase I test database (all video sequences included). The vertical and horizontal axes represent the subjective and the objective scores, respectively. Each sample point represents one test video sequence. (a) PSNR; (b) PSNR with proposed weighting method; (c) SSIM; (D) SSIM with proposed weighting method. All SSIM values were raised to the 8th power for visualization purpose only.

might be better ways to combine the information content and the perceptual uncertainty measures. Second, the computation of local image contrast and the estimation of motion vectors may be improved. For example, we frequently observe instabilities in the current optical flow-based motion estimation algorithm, especially in the video frames with large background motion. This implies that more robust motion estimation method is needed in the existence of noise and large motion. Third, the sophistication and the high-level nature of the proposed model make its parameters difficult to calibrate. More careful psychovisual studies are still needed. Fourth, the weighting function computed based on our model is effective and consistent in improving the performance of VQA algorithms in all the tests we have done so far (with MSE/PSNR and SSIM). Other VQA algorithms may also be included to further validate the model. Finally, the general idea of quality map weighting does not constrain itself to be used for full-reference VQA only, as being tested in this paper (Note that both the MSE/PSNR and the SSIM calculations require access to the original video sequence as a reference). If a no-reference method is available that can provide us with a quality map without using any reference video, the same weighting approach is also applicable. Such no-reference or blind VQA systems are highly desirable in the real world and are yet to be developed in the future.

6. Acknowledgement

We would like to thank Prof. Eero P. Simoncelli, Dr. Alan A. Stocker, and the anonymous reviewers for their constructive comments and Prof. Michael J. Black for kindly providing the program for optical flow estimation.

References

1. T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, A. Bovik, ed. (Academic Press, 2nd Ed.), 923–939 (2005).
2. Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, Eds. (CRC Press), 1041–1078 (2003).
3. B. A. Wandell, *Foundations of Vision*. (Sinauer Associates, Inc.), (1995).
4. H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessement," *The First Inter. Workshop on Video Proc. and Quality Metrics for Consumer Electronics*, Jan (2005).
5. Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Proc.: Image Comm.*, **19**, 121–132 (2004).

6. Z. Wang, H. R. Sheikh, A. C. Bovik and E. P. Simoncelli, "Image quality assessment: from error visibility to structual similarity," *IEEE Trans. Image Processing*, **13**, 600–612 (2004).
7. Z. K. Lu, W. Lin, X. K. Yang, E. P. Ong, and S. S. Yao, "Modeling visual attention's modulatory afterreffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, **14**, 1928–1942 (2005).
8. K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," *IEEE Inter. Conf. Acoustics, Speech, and Signal Processing*, April (2007).
9. K. Seshadrinathan and A. C. Bovik, "An information theoretic video quality metric based on motion models," *Third Inter. Workshop on Video Proc. and Quality Metrics for Consumer Electronics*, Jan (2007).
10. Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," *IEEE Inter. Conf. Acoustics, Speech, and Signal Proc.*, 573–576 (2005).
11. H. R. Sheikh, A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing*, **15**, 430–444 (2006).
12. A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, **9**, 578–585 (2006).
13. E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," *IEEE Inter. Conf. Computer Vision and Pattern Recognition*, 310–315 (1991).
14. Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts," *Nature Neuroscience*, **5**, 598–604 (2002).
15. F. Hürlimann, D. C. Kiper, and M. Carandini, "Testing the Bayesian model of perceived speed," *Vision Research*, **42**, 2253–2257 (2002).
16. E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, **24**, 1193–1216 (2001).
17. R. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik, "Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy," *J. Opt. Soc. Am. A*, **22**, 2039–2049 (2005).
18. J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, **434**, 387–391 (2005).
19. Z. Wang and X. L. Shang, "Spatial pooling strategies for perceptual image quality assessment," *Proc. IEEE Inter. Conf. Image Processing*, (2006).
20. Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, **12**, 243–254 (2003).
21. VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," http://www.vqeg.org/, (2000).

22. Black, M. J. and Anandan, P., "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, **63**, 75–104 (1996).
23. T.Vlachos, "Simple method for estimation of global motion paramenters using spares translational miton vector fields," *Electronics letters*, **34**, 90–91 (1998).
24. Peli, E., "Contrast in complex images," *J. Opt. Soc. Am. A*, **7**, 2032–2040 (1990).
25. Teo, P. C. and Heeger, D. J., "Perceptual image distortion," *Proc. SPIE*, **2179**, 127–141 (1994).
26. Heeger, D. J. and Teo T. C., "A model of perceptual image fidelity," *IEEE Inter. Conf. Image Processing*, 343–345 (1995).