Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation

Zhongkang Lu, *Senior Member, IEEE*, Weisi Lin, *Senior Member, IEEE*, Xiaokang Yang, *Senior Member, IEEE*, EePing Ong, and Susu Yao

***Abstract*—With the fast development of visual noise-shaping related applications (visual compression, error resilience, watermarking, encryption, and display), there is an increasingly significant demand on incorporating perceptual characteristics into these applications for improved performance. In this paper, a very important mechanism of the human brain, visual attention, is introduced for visual sensitivity and visual quality evaluation. Based upon the analysis, a new numerical measure for visual attention's modulatory aftereffects, perceptual quality significance map (PQSM), is proposed. To a certain extent, the PQSM reflects the processing ability of the human brain on local visual contents statistically. The PQSM is generated with the integration of local perceptual stimuli from color contrast, texture contrast, motion, as well as cognitive features (skin color and face in this study). Experimental results with subjective viewing demonstrate the performance improvement on two PQSM-modulated visual sensitivity models and two PQSM-based visual quality metrics.**

***Index Terms*—Just-noticeable difference (JND), noise shaping, perceptual quality significance map (PQSM), visual attention, visual quality evaluation, visual sensitivity.**

I. Introduction

With THE fast development of visual noise-shaping related applications (e.g., visual compression, error resilience, watermarking, encryption, and display), research on human visual sensitivity analysis and quality evaluation has drawn a lot of attention from scientists and researchers [1]–[7]. Visual sensitivity refers to the ability of human observers to detect noise or distortion in the view field. Numerically, visual sensitivity can be regarded as the inverse of the just-noticeable difference (JND), which determines the visibility thresholds in pixels [8], [9] or subbands [5], [10]. Visual quality metrics (VQMs) are designed to predict perceived image and video quality by measuring the detectability [4], [5], [11]–[14] or annoyance of noise/distortion [2], [15]–[17] introduced via visual processing.

Visual noise shaping is to allocate the inevitable noise or distortion into some subbands or spatial areas so that the resultant visual variation is the least noticeable or annoying to the human visual system (HVS). With regard to the human visual perception, the noise or distortion introduced into image/video can be classified into three categories: 1) imperceivable noise; 2) near-threshold noise; and 3) suprathreshold distortion. Imperceivable noise is below JND and, therefore, hard to be perceived by the HVS. Near-threshold noise is just above the thresholds while suprathreshold distortion is much stronger than JND. Generally, suprathreshold distortion appears in the form of structural patterns, such as blockiness, ringing, blurring and jerkiness in decoded visual signal [18].

Visual attention is the result of several millions of years of evolution [19], and the research on visual attention began more than 100 years ago [20]. It can be defined as a set of strategies to reduce the computational cost of the search processes inherent in visual perception [21]. It has *top-down* (or knowledge/task-driven) and *bottom-up* (or stimulus-driven) mechanisms [22]. In the former mechanism, attention is under the overt control of the subject and related to cognition processing in the human brain [23], [24]; it is voluntary, effortful, and has a slow (sustained) time course [25]. In the latter mechanism, attention is driven by external stimuli and some fast perception processing of the human brain draws attention to a particular location; it is automatic and has a transient time course. Generally, the stimuli involved in *bottom-up* control include luminance, color, orientation and motion contrast, while the features involved in *top-down* control are pattern, shape, and other cognitive processing related features. Moreover, audition, touching, and other sensories also affect visual attention [26]. Because the *top-down* control has a much longer time course, it may play a more important role on the shift and distribution of visual attention. Visual attention modulates all levels of visual perception [27], including visual sensitivity [28] and, therefore, visual quality evaluation.

The simplified concept of visual attention has been adapted for video quality evaluation [29], [30]. In [29], under a simple assumption that the HVS' focus position is on the center of the image, a weighted SNR metric is proposed according to the eccentricity and contrast sensitivity function (CSF), based on a fixed gradient model of visual attention. In [30], a *bottom-up* visual attention model is proposed to weight the visual quality metric [11] for accuracy improvement, mainly based on location, contrast, color, luminance, and motion, without consideration of a gradient model. There is a need for automatic estimation to include both *bottom-up* stimuli and *top-down* features for visual sensitivity and quality evaluation. Moreover, the motion suppression effect [32]–[34] has to be considered since motion affects visual sensitivity and quality assessment significantly in

Manuscript received March 31, 2004; revised October 5, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhigang (Zeke) Fan.

Z. Lu, W. Lin, E. Ong, and S. Yao are with the Institute for Infocomm Research, Agency for Science, Technology, and Research, Singapore 119613 (e-mail: zklu@i2r.a-star.edu.sg; wslin@i2r.a-star.edu.sg; epong@i2r.a-star.edu.sg; ssyao@i2r.a-star.edu.sg).

X. Yang is with the Department of Electronic Engineering, Institute of Image Communication and Information Processing, Shanghai Jiaotong University, Shanghai 200030, China (e-mail: xkyang@ieee.org).

Digital Object Identifier 10.1109/TIP.2005.854478

video; integration of multiple stimuli and features would allow the application to a wider scope of visual signal; and a flexible gradient model is more realistic in the human perception [35], [36].

In this paper, perceptual quality significance map (PQSM) is proposed to reflect the modulatory aftereffects of visual attention, on visual sensitivity and quality evaluation. The PQSM is automatically estimated with both *bottom-up* stimuli and *top-down* features. A general formulation is proposed for multiple stimuli/feature integration to capture the basic ideas of visual attention for practical applications. Color contrast, texture contrast, motion, skin color, and face features are extracted and integrated in the current implementation, under a flexible gradient model. The PQSM for an image or video can be incorporated in JND estimators and VQMs enhanced performance.

The rest of this paper is organized as follows. Section II reviews the related work on biological and psychological mechanisms of visual sensitivity and visual attention, as well as on existing models for JND estimation, VQMs and visual attention. The proposed computational model of PQSM generation, two PQSM-modulated JND models, and two PQSM-based VQMs are presented in Sections III–V, respectively. The experimental results are demonstrated in Section VI, in comparison with the associated subjective test data. The conclusion and future work are given in Section VII.

II. Related Research Work

This section reviews the previous work related to the proposed PQSM models. In Section II-A, biological and psychological evidence is presented for visual sensitivity, followed by a review of the existing JND estimators. Section II-B introduces the current VQMs. In Section II-C, characteristics of and relevant research on visual attention are discussed; the section also provides the ground of some strategies for the proposed PQSM estimation algorithm in Section III. In Section II-D, current techniques combining visual attention with various visual processing tasks are introduced.

A. Visual Sensitivity Analysis

1) Biological and Psychological Mechanisms Behind Visual Sensitivity: In general, visual sensitivity includes amplitude sensitivity, motion sensitivity [37], and flicker sensitivity [38]. Among them, amplitude sensitivity is the basic. The latter two are closely linked, and much more complex than amplitude sensitivity. Amplitude sensitivity has been intensively explored, and is typically modeled as the output of the biological and psychological mechanisms of the HVS [7], [39], [40].

Visual sensitivity results from the anatomy, the limitations and imperfections of the human eye, such as the optical properties of the HVS, the photo-electric transmission curve of photoreceptor, the distribution of photoreceptor on retina, the response of bipolar cells and ganglion cells in the second layer of retina, and the noise introduced in signal via vision path. After entering the human brain, visual signal is mainly processed in cortex area V1. Electro-physiological experiments have shown that the response of neurons in V1 exhibits a band limited properties [41]. The HVS decomposes visual data into perceptual channels with spatial/temporal frequencies, orientations, and colors [42].

Masking between two or among more visual channels usually results in an increase of the visibility thresholds [43], [44]. The masking in zero frequency channel is called luminance adaptation, the masking in nonzero frequency channel is called contrast masking, and the masking within a channel or with other channel(s) is called intrachannel masking and interchannel masking. The common factors considered for visual visibility thresholds are: 1) spatial and temporal frequencies, 2) luminance, and 3) contrast orientations.

The visual sensitivity is enhanced on precued spatial locations [45]–[47], due to the aftereffects of visual attention. The visibility thresholds in the precued areas are lower than the other nonattentional areas, with the recorded differences varying from 0 to 9.4 dB because of different experimental designations. Itti *et al.* reported that visual attention can elevate the sensitivity with spatial and temporal frequencies by 30%, the sensitivity with orientations by 40%, and the sensitivity peak altitude by 5.2 dB [28].

Another global factor affecting visual sensitivity is motion suppression [32], [33] caused by the motion of object projection on retina. It shows the suppression can reach about 0.6 log units, or 12 dB in maximum. It is believed that motion on retina image increases the processing cost of visual perception and, therefore, suppresses the visual sensitivity. Motion suppression happens in the low attentional areas when the motion is different to that in high-attentional areas. Since the eye movement follows the shift of visual attention, motion suppression can be regarded as another aftereffect of visual attention. It is worth noting that inaccurate pursuit of eye movement also brings about motion on retina in saccadic condition, which causes motion suppression.

2) Computational Models for Visual Sensitivity: The existing computational visual sensitivity or JND models can be classified into two categories: pixel based [8], [9] and transform based [5], [10], [48], [49] (extensively investigated especially in DCT domain). A subband domain JND estimator can be more precise, because it can really take into account the different interaction between signals or components in the masking effects. However, the pixel domain estimation is computationally simpler, and has its advantages in some applications, e.g., motion estimation (before subband coefficients are available), perceptual evaluation for already-decoded images/video.

In Chou and Li's model [8], the JND of a pixel (x, y) is obtained as

$$\mathrm{JND} = \max\{f_1(bg, mg), f_2(bg)\} \tag{1}$$

where bg is the average background luminance, mg is the contrast value, which is the maximum output of high-pass filtering at four directions, and $f_1(bg, mg))$ and $f_2(bg))$ represent contrast masking and luminance adaptation, respectively (for 8-bit image presentation)

$$f_1(bg, mg) = mg \cdot \alpha(bg) + \beta(bg) \tag{2}$$

$$f_2(bg) = \begin{cases} T_0 \cdot \left(1 - \left(\frac{bg}{127}\right)^{0.5}\right) + 3, & \text{if } bg \le 127 \\ \gamma \cdot (bg - 127) + 3, & \text{if } bg > 127 \end{cases} \tag{3}$$

where $\alpha(bg) = bg \cdot 0.0001 + 0.115$ and $\beta(bg) = \gamma - bg \cdot 0.01$. All the parameters were empirically determined by fitting the model with subjective test results [8] under certain viewing conditions (i.e., a monitor with its associated gamma function, viewing distance, ambiance illumination). The conversion between grey levels and display luminance has been factored in the valuation of the parameters in the above equations. The model is only for luminance components.

Yang *et al.* [9] extended Chou and Li's model to account for multiple channels and the combined effect of contrast masking and luminance adaptation, and, therefore, the spatial JND threshold for a pixel can be expressed as

$$\Theta_s^i = \Theta_z^i + \Theta_c^i - C_{zc}^i \cdot \min\left(\Theta_z^i, \Theta_c^i\right) \tag{4}$$

where Θ_z represents the luminance adaptation in each zero frequency channel, Θ_c^i represents the intrachannel contrast masking, C_{zc}^i reflects the interchannel masking between zero frequency channel and contrast channel, and i denotes Y, Cb and Cr in $Y - Cb - Cr$ space. Combined with temporal masking, the final JND is obtained as

$$\Theta^i = \Theta_s^i \cdot \Theta_t^i \tag{5}$$

where Θ_t^i is the recorrection function for temporal masking [50].

Watson *et al.* proposed an analytic formula for JND thresholds on each DCT frequency component [5] as the product of a luminance adaptation T_0, a temporal contrast masking function $T_w(w)$, a spatial contrast masking function $T_f(u,v)$, and an orientation suppression function $T_a(u,v)$

$$T(u,v,w) = T_0 \cdot T_w(w) \cdot T_f(u,v) \cdot T_a(u,v) \tag{6}$$

where u, v, and w represent spatial horizontal frequency, spatial vertical frequency, and temporal frequency, respectively. $T_w(w)$ is the inverse of the magnitude response of a first-order discrete IIR low-pass filter with a sample rate of w_s and a time constant of τ_0

$$T_w(w) = \left| \frac{e^{\frac{1+i2\pi\tau_0 w}{\tau_0 w_s}} - 1}{e^{\frac{1}{\tau_0 w_s}} - 1} \right|. \tag{7}$$

$T_f(u,v)$ is the inverse of a Gaussian function

$$T_f(u,v) = \exp\left(\pi \frac{u^2+v^2}{f_0^2} \left(\frac{p}{16}\right)^2\right) \tag{8}$$

where p is the display resolution in pixels/degree, and the factor of $p/16$ converts from DCT frequencies to cycles/degree; f_0 corresponds to the radial frequency at which the threshold is elevated by a factor of e^{π}. $T_a(u,v)$ is expressed as

$$T_a(u,v) = \frac{2^{\frac{\zeta-1}{\zeta}}}{1 - \frac{4\eta u^2 v^2}{u^2+v^2}} \tag{9}$$

where η and ζ are parameters. All parameters were decided by fitting with the subjective test results.

B. Visual Quality Metrics (VQMs)

Visual sensitivity plays an important role in visual quality gauging. Imperceivable noise has no or relatively insignificant contribution on the change of visual quality. Under near-threshold conditions, visual quality is evaluated by measuring the detectability of distortion; under suprathreshold conditions, visual quality is evaluated by measuring the annoyance of distortion. In the visual communication applications, imperceivable noise usually results from perceptually lossless compression; near-threshold noise usually results from lossy compression at medium or high bit rates; and suprathreshold distortions exist in low or very low bit-rate compression. Measuring the annoyance of suprathreshold distortions is very complex, since it is a highly subjective process that involves with the high level activities of human brains, such as pattern matching, object recognition and scene perception. It depends on both the distortion characteristics and the original visual contents, and artifacts differing qualitatively in their appearance may produce different levels of annoyance even though they have the same sensitivity threshold and the same error energy [51].

Some VQMs measure the detectability of distortion [4], [5], [11]–[14]. Among these VQMs, Teo *et al.* [11] and Winkler [4] used steerable pyramid transform to separate input video into several channels, and the contrast gain control has been realized by an excitatory nonlinearity function. JNDs can be directly used to facilitate the distortion evaluation, like in Watson's metric [5] in DCT domain and Lubin's metric [12] in spatial domain. In [14], perceptual errors are evaluated in flat, edge and texture regions, respectively, before integration by Fuzzy integral. Le Callet *et al.* [13] proposed another color image quality metric; based on their psychophysical experimental results, the visual representations of errors distributed over color, spatial, and frequency dimensions between two images are computed and then evaluated via error pooling.

Some other VQMs measure the strength of structural distortions. In [2] and [15]–[17], prior knowledge of coding (structural) artifacts specific for decoded images/video is used in VQMs. Blockiness, ringing, and blurring are oft-occurring coding artifacts [18]. Karunasekera's model [2] estimates blockiness artifacts via horizontal and vertical high-pass filters and masked edge errors via a nonlinear transform. A no-reference blocking artifact measurement algorithm proposed by Wu [15] uses a weighted mean square difference along block boundaries as the blockiness measure. In the structural similarity index (SSIM) proposed by Wang *et al.* [16], visual distortion is evaluated with luminance, contrast, and structural changes, as well as motion in video. Ong *et al.* [17] detected and combined three major disturbing artifacts (namely, damaged edge, blockiness, and ringing) to give final quality scores for low bit-rate visual communication.

C. Visual Attention Models

Biological research has proved that allocating attention to a spatial location in the visual field is associated with an increase in cortical response at that location [52], and this means more HVS computational resource is allocated to high attentional areas than low attentional areas. The formation of visual attention is a very complex process concerning all aspects of visual processing in the human brain, such as processing of visual

contents [19], stimuli and feature integration [53], [54], feature binding [55], and object perception.

Moreover, visual attention has capacity limit [56], in spite of some arguments in this topic [57]. Perceptual processing of multiple items or large attentional areas are not independent and the visual sensitivity enhancement can be reduced in such cases, due to the limited computational resource of the human brain.

The shape of high attentional area can be arbitrary [58], and the gradient of visual attention is flexible [35], [36]. Two special cases are the *attentional gradient* model [59] and the *zoom-lens* model [60]. In the vicinity of a highly attentional object, such as a human head, the gradient of attention is quite steep [61], and sometimes the distribution can be discrete—as in the *zoom lens* model.

Other properties of visual attention are listed as follows.

- The HVS is not blind out of attention, although some research shows that some big and flashy changes could be ignored in peripheral vision [62], [63] (due to the fact that the limited capacity of the human brain, especially short-term visual memory [64], [65], may be so engaged with the current visual objects that no resource can be reallocated to process new changes).
- Visual attention enhances not only the visual sensitivity, but also observers' performance in a wide variety of other visual tasks, e.g., the integration of multiple stimuli [53], [54].
- Visual attention can be either object based or area based [66].
- The integration of multiple stimuli on visual attention is nonlinear additivity [67].
- Most importantly, it is stimulus contrast rather than absolute stimulus strength that guides the bottom-up attention [67]–[69].

Since Broadbent [71] first proposed the filter theory of selective attention, a number of computational models on visual attention have been developed [72]–[75] based on Treisman's stimulus integration theory [76]. All these models adapt a two-stage framework: The first stage preattentively processes all incoming visual information equally and in a parallel fashion; the second stage filters and combines the extracted information to form a salience map. The final attentional position is selected by a embedded decision module.

D. Combining Visual Attention With Visual Processing

The concept of visual attention has been adopted in visual processing in somewhat simplified forms [29], [77]–[83]. In [77] and [78], region of interest (ROI) can be regarded as a discrete visual attentional map. The maximum shift [79] method and the generalized bitplane-by-bitplane shift method [84] were proposed to enhance the ROI coding quality. In [80], Reddy proposed a *level of details* control algorithm in virtual environment to remove extraneous details which the user cannot perceive and, thus, to optimize the computational resource assignment on rendering and display with little or no perceptual artifacts. Wang *et al.* [81] gave a foveation scalable video coding algorithm for image compression with preselected fixation points. A similar technique based on a fixed gradient model of visual attention was used by Lee *et al.* [29], [82], [83] to reduce spatial resolution of image nonuniformly, and the resolution reduced image can be then used for optimal rate control in video compression [82], image quality assessment [29], and video communication [83]; the eye fixation positions are either predefined or found by an eye tracker.

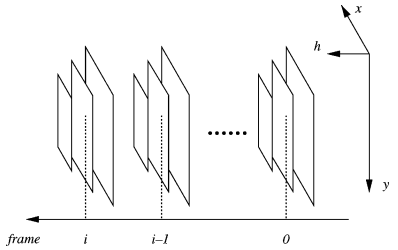


Fig. 1. General hierarchical PQSM [h: hierarchical maps for an image (from the full-resolution map to the roughest map)].

Some other techniques with embedded visual attention estimation were proposed in [30], [85]–[87]. As mentioned in the Introduction, Osberger *et al.* [30] combined his *bottom-up* visual attention model with Teo's VQM [11] for accuracy improvement by adding weights to attentional areas. In the multiple resolution rendering technique proposed by Cater *et al.* [85], a task map built by a *bottom-up* visual attentional model modulates the spatiotemporal CSF to guide a progressive animation system taking full advantage of image-based rendering. In Dhavale *et al.*'s paper [86], Itti's attention model [74] is combined with a foveation filter to keep more details in video at predicted eye fixation positions. Yang *et al.* [87] used a visual sensitivity map modulated by skin color based attention for rate control in videophone compression applications.

III. Perceptual Quality Significance Map (PQSM) Estimation

As we have already known, more computational resource of the human brain is allocated to high attentional areas than low attentional areas, and this is the reason of visual attention's modulation on visual sensitivity in different areas. We propose a new numerical expression, PQSM, to represent the combined effect of the modulation of visual attention and the extra computational cost imposed by motion suppression, on visual sensitivity and visual perception. The PQSM is designed to reflect the statistical allocation of the human brain's processing resource on local visual contents. We do not attempt to implement a full visual attention model but rather aim at a practical solution inspired by the relevant physiological and psychological evidence. The concept was first proposed in [88], and improved in [89]. As pointed out in Section II-A1, conceptually, the influence of motion suppression can be regarded as an aftereffect of visual attention. However, for the convenience in computation, the PQSM for an image can be determined as the product of the influence of motion suppression and the visual attention measure derived from the other stimuli/features, since motion suppression is a global factor (Section III-C). The hierarchical PQSM can be extended to the temporal axis for video, as illustrated in Fig. 1.

The proposed three modules to generate a PQSM are illustrated in the left-hand side portion of Fig. 2: 1) the visual

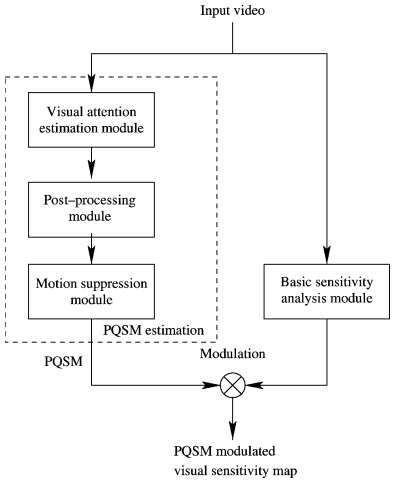


Fig. 2. Flowchart of generating PQSM modulated visual sensitivity map.

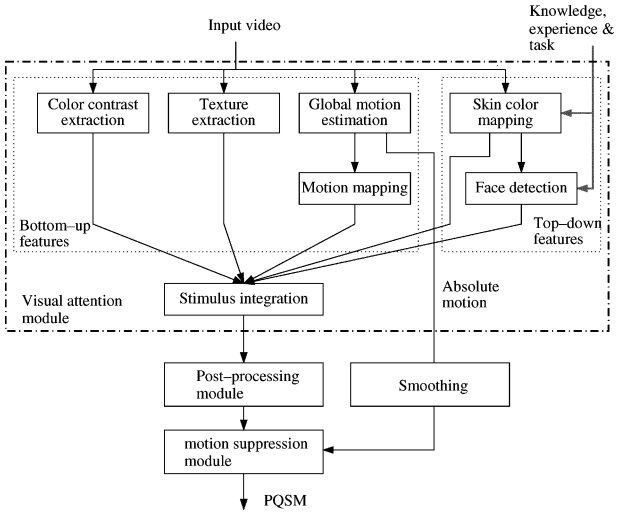


Fig. 3. Flowchart of PQSM generation.

attention module to extract and integrate multiple stimuli/features to form a visual attentional map; 2) the post-processing module to smooth and implement a flexible gradient model for the modified visual attentional map; and 3) the motion suppression module to include motion influence for the final PQSM. Fig. 3 shows the next level of details for the PQSM generation. Only those well-established findings on visual attention presented in Section II will be modeled in the proposed algorithm.

A. Visual Attention Module

The proposed visual attention model is driven by both *bottom-up*, space-based stimuli (color, texture, and motion) and *top-down*, object-based features (skin color and face). Selection of the stimuli/features is based on a balance of computational efficiency and output accuracy.

1) Derivation of Stimuli and Features:

- *Color contrast*: Color contrast is one of the basic stimuli that draws attention, and a bigger color difference from the background color usually attracts higher attention. A three-step approach has been developed for deriving and scaling a color contrast stimulus in an RGB color image. 1) The dynamic k-means algorithm clusters the 8×8 image regions with a predefined variation. If the size of the biggest cluster is more than a threshold (e.g., a half of the image), the background color (R_c, G_c, B_c) is calculated as the mean in the cluster; otherwise, color contrast is not a stimulus in the image, because only pixels with color sufficiently distinguished from the background attract attention (if no obvious *reference* cluster exists, no pixel would stand out due to its color). 2) The distance between a color value and the background color is calculated by

$$\mathrm{dis} = \sqrt{(R-R_c)^2 + (G-G_c)^2 + (B-B_c)^2}. \tag{10}$$

The distance histogram is constructed: $\mathrm{His} = \{n_{\mathrm{his}-d}\}$, and $n_{\mathrm{his}-d}$ is the number of pixels in the image with $\mathrm{dis} = d$. 3) A scale is calculated by

$$\mathrm{scale} = \frac{\sum\limits_{d} n_{\mathrm{his-d}}}{\sum\limits_{d} n_{\mathrm{his-d}} * d}. \tag{11}$$

The color contrast stimulus is scaled as

$$s_c = \mathrm{scale} * \mathrm{dis}. \tag{12}$$

Finally, s_c is truncated so that $s_c \in [0, s_c^{\max}]$, where $s_c^{\max} = 0.5$ is used in this paper. Obviously, the bigger color variation from the dominant color a pixel is, the bigger contribution it has toward the PQSM. If more pixels in an image have bigger color variation, the impact of a certain color variation will not be as significant as the situation otherwise.

- *Texture contrast*: The average gradient is obtained with horizontal and vertical Sobel operators in a region (of size 3×3 in this paper). The texture stimulus s_t is derived and scaled with a similar three-step method as in the formation of color contrast stimulus.
- *Motion*: Motion is one of the major stimuli on visual attention [90], [91]. Motion detected from video can be divided into relative motion and absolute motion. The former is the object motion against the background or other objects in the scene, while the latter is the motion against the frame of viewing (the combination of camera motion and object motion in the real-word coordinates[1]). Relative motion is a stimulus on visual attention.

Black's multiple layer dense flow estimation algorithm [92] is adopted to estimate the absolute motion v_a, and the relative motion v_r is estimated via Zhang's estimation algorithm [93]. The scaled relative motion v_r' is obtained with the similar scaling procedures as Steps 1)–3) in the estimation of color contrast stimulus; Step 1) for clustering is not needed because the relative motion of background is always zero.

It is not easy to evaluate the effect of v_a and v_r toward visual significance in general. Based upon the observation on most digital images in practical use, here we give a simple heuristic set of rules. 1) Usually, the

[1]If the camera follows exactly the object, the absolute motion is zero.

TABLE I
MOTION ATTENTIONAL LEVEL BY RELATIVE MOTION AND ABSOLUTE MOTION

relative motion	absolute motion	motion attentional level
low	low	low
low	high	low
high	low	high
high	high	moderate

attentional level of an object is low when relative motion is low (as shown in the upper two rows of Table I). 1) The attentional level of an object is relatively significant when relative motion is high. However, since the camera's motion often indicates the most important object/region in the visual field, the attentional contribution of a motion stimulus for an object with both high absolute motion and high relative motion is merely moderate (because the object is usually not of the primary interest); the highest attentional contribution occurs with low absolute motion and high relative motion. These two circumstances are represented as the two lower rows in Table I.

The motion stimulus generated by relative motion should be adjusted by absolute motion (i.e., motion mapping)

$$s_m = v_r' \cdot g_{\text{adj}}(v_r', v_a) \tag{13}$$

where $g_{\text{adj}}()$ is an adjusting function defined by Table II, based on the concept of Table I and the experimental results with the standard video sequences commonly in use.

- *Skin color*: We include some useful cognitive features to achieve a more effective extraction for practical usage. Human body and warm color [74] are always cognitive-related features on visual attention. A statistical model is adopted to detect regions with skin color on $Cb - Cr$ domain. The skin color stimulus is denoted as s_s.
- *Face*: As another cognitive feature, face is a more significant stimulus on visual attention. Rowley's face detection algorithm [94] is used to locate faces in image. The result of skin color detection is used to recorrect the face detection result because Rowley's algorithm is merely based on gray-scale image. The face stimulus is denoted as s_f.

2) Integration: Although it has been proven that *bottom-up* controlled attention and *top-down* controlled attention are processed in different areas in the brain [36], [95], their integration is mostly processed in the prefrontal cortex [95]. Nothdurft's nonlinear additivity model [67] is adapted to integrate the stimuli and features together. In his model, the combined saliency effect of stimuli s_1 and s_2 can be expressed as

$$s_{12} = s_1 + s_2 - s_{12}^* \tag{14}$$

where

$$s_{12}^* = \min(c_{12} \cdot s_1, c_{21} \cdot s_2) \tag{15}$$

where s_{12}^* represents the adjustment for the combined effect, c_{12} and c_{21} represent the cross-dimensional coupling factors between s_1 and s_2, and c_{12} and c_{21} fall in the range of [0, 1].

In this paper, (14) is extended to N stimuli/features

$$S_N = \sum_i^N s_i - \sum_i^N f(c_{ip} \cdot s_p, c_{pi} \cdot s_i) \tag{16}$$

where $p = \arg\max_i(s_i)$, $i = 1, \cdots, N$; $f(\cdot,\cdot)$ is an appropriate nonlinear function, and, in this paper, $f(\cdot,\cdot) = \min(\cdot,\cdot)$ is chosen; only the coupling with the main stimulus/feature is considered. A bigger value for $c_{f,s}$ ($c_{f,s} = 0.75$ in the current implementation) is adopted because of the high correlation between the face and skin color. On the other hand, it is believed that color and luminance contrasts attract independent attention [70], so we have $c_{cl,ct} = 0$. We have $c_{m,f} = c_{m,s} = 0.5$, and the other coupling factors are set to 0.25. Obviously, more research is needed in determining the optimum parameters.

Equation (16) satisfies

$$S_{i-1} \le S_i \le S_{i-1} + s_i \tag{17}$$

where S_{i-1} represents the combined effect of $i-1$ visual stimuli. We can see that (14) is equivalent to (16) when $N = 2$.

B. Post-Processing Module

For efficiency, the post-processing module is performed on block representation

$$\text{vam}_1(a,b) = \frac{\sum_{(x',y') \in \Re_{a,b}} S_N(x',y')}{L} \tag{18}$$

where (a,b) is index of block representation, $\Re_{a,b}$ is a collection of pixels in block (a,b), L is the size of $\Re$, and $S_N(x',y')$ are visual attention values in block (a,b). In this paper, the block size is set to 8×8 and $L = 64$.

A flexible kernel is used to mimic the flexible gradients with visual attention [35], [36] and is defined as

$$\ker_0(a-a', b-b') = \begin{cases} 1, & \text{if } \rho \le \sigma \\ e^{-\frac{(\rho-\sigma)^2}{\sigma^2}}, & \text{if } \rho > \sigma \end{cases} \tag{19}$$

where $\rho = \sqrt{(a-a')^2 + (b-b')^2}$ and $\sigma = \text{vam}_1(a,b) + 1$. The normalized kernel is obtained by

$$\ker(a-a', b-b') = \frac{\ker_0(a-a', b-b')}{\sum_{a',b'} \ker_0(a-a', b-b')}. \tag{20}$$

Obviously, the proposed kernel fits the flexible gradients of visual attention: When the value of $\text{vam}_1(a,b)$ is high, the result of the kernel is steep, and vice verse.

The enhanced visual attentional map is derived as

$$\text{vam}_2(a,b) = \max_{a',b'} \left(\text{vam}_1(a',b') \cdot \ker(a-a', b-b')\right). \tag{21}$$

TABLE II
ADJUSTING FUNCTION $g_{\mathrm{adj}}(v'_r, v_a)$ ($v'^{\max}_r$ IS THE MAXIMUM VALUE OF SCALED RELATIVE MOTION)

	$v_a \in [0.0, 1.0)$	$[1.0, 2.0)$	$[2.0, 3.0)$	$[3.0, 4.0)$	$[4.0, 5.0)$	$[5.0, \infty)$
$v'_r \in [0.0, \frac{1}{3}v'^{max}_r)$	1.0	1.0	1.0	1.0	1.0	1.0
$[\frac{1}{3}v'^{max}_r, \frac{2}{3}v'^{max}_r)$	1.0	1.0	0.9	0.9	0.8	0.8
$[\frac{2}{3}v'^{max}_r, v'^{max}_r)$	1.0	0.9	0.8	0.7	0.6	0.5

TABLE III
SUPPRESSION ON VISUAL ATTENTION BY ABSOLUTE MOTION

visual attentional level (vam_3)	absolute motion (v_a)	motion suppression
low	low	weak
low	high	strong
high	low	weak
high	high	moderate

TABLE IV
ABSOLUTE MOTION SUPPRESSION FUNCTION $f_{\mathrm{ms}}(\mathrm{vam}_3, v_a)$

	$v_a \in [0.0, 3.0]$	$[3.0, 4.0)$	$[4.0, 5.0)$	$[5.0, 6.0)$	$[6.0, 7.0)$	$[7.0, 8.0)$	$[8.0, \infty)$
$vam_3 \in [0, 1/3) \cdot vam_3^{max}$	1.0	0.9	0.8	0.7	0.7	0.6	0.6
$[1/3, 2/3) \cdot vam_3^{max}$	1.0	1.0	1.0	0.9	0.8	0.8	0.7
$[2/3, 1.0) \cdot vam_3^{max}$	1.0	1.0	1.0	1.0	1.0	0.9	0.8

To address the visual attention's capability limit issue mentioned in Section II-C, the visual attentional map is further modified as

$$\mathrm{vam}_3(a,b) = \begin{cases} \mathrm{vam}_2(a,b), & \text{if } \sum_{a,b} \mathrm{vam}_2(a,b) \le \mathrm{VA} \\ \frac{\mathrm{vam}_2(a,b) * \mathrm{VA}}{\sum_{a,b} \mathrm{vam}_2(a,b)}, & \text{if } \sum_{a,b} \mathrm{vam}_2(a,b) > \mathrm{VA} \end{cases} \tag{22}$$

where $\mathrm{VA} = 1.0 \times N_a \times N_b$ is a predefined constant to reflect the visual attentional capacity of the human brain. $N_a \times N_b$ are the horizontal and vertical block size of video.

C. Motion Suppression Module

Motion suppression is caused by object motion on the retina, which can be measured by absolute motion v_a if the eye's movement on the visual field is smooth and slow. The relationship between visual attention and motion suppression is outlined in Table III, so the motion suppression effect can be determined as the product of vam_3 and the influence of motion suppression $f_{\mathrm{ms}}()$

$$\mathrm{pqsm} = \mathrm{vam}_3 \cdot f_{\mathrm{ms}}(\mathrm{vam}_3, v_a) \tag{23}$$

where $0 < f_{\mathrm{ms}}() \le 1$ is a function defined in Table IV, which is an implementation of Table III. Table IV is set for $v_a \le 8$ pixels, in line with the response speed of actual video acquisition devices (if $v_a > 8$ pixels, the image is prone to blurring error). Tables II and III are defined for video in PAL (720 × 576, 50 Hz) format. As for NSTC (720 × 486, 60 Hz) video, v_a is adjusted by multiplying by a factor of 6/5. In general, the motion measures in (23) should be degrees per second. With fixed viewing distance and display framerate, pixels per frame is equivalent to degrees per second.

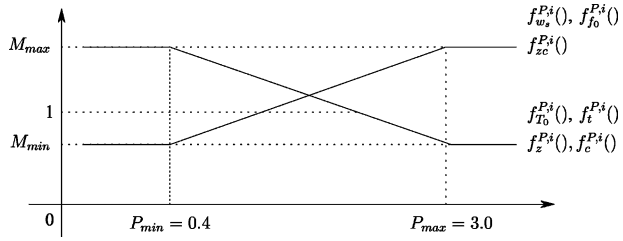


Fig. 4. Modulation functions for PQSM-modulated JND Model 1 and 2.

IV. PQSM-MODULATED VISUAL SENSITIVITY MODELS

Fig. 2 shows how the PQSM modulates a visual sensitivity model. Yang's JND model [9] and Watson's JND model [5] will be demonstrated with PQSM modulation, and the resultant modulated models are hereinafter referred to as PQSM-modulated JND Model 1 and PQSM-modulated JND Model 2, respectively.

PQSM-modulated JND Model 1 can be expressed as

$$\Theta^{P,i}_s = \Theta^{P,i}_z + \Theta^{P,i}_c - C^{P,i}_{zc} \cdot \min\left(\Theta^{P,i}_z, \Theta^{P,i}_c\right) \tag{24}$$

$$\Theta^{P,i} = \Theta^{P,i}_s \cdot \Theta^{P,i}_t \tag{25}$$

where $\Theta^{P,i}$, $\Theta^{P,i}_s$, $\Theta^{P,i}_t$, $\Theta^{P,i}_z$, $\Theta^{P,i}_c$, and $C^{P,i}_{lc}$ denote the modulated versions of the variables defined in (4) and (5), and

$$\Theta^{P,i}_z = \Theta^i_l \cdot f^{P,i}_z(\mathrm{pqsm}) \tag{26}$$

$$\Theta^{P,i}_c = \Theta^i_c \cdot f^{P,i}_c(\mathrm{pqsm}) \tag{27}$$

$$C^{P,i}_{zc} = C^i_{zc} \cdot f^{P,i}_{zc}(\mathrm{pqsm}) \tag{28}$$

$$\Theta^{P,i}_t = \Theta^i_t \cdot f^{P,i}_t(\mathrm{pqsm}) \tag{29}$$

where $f^{P,i}_t()$, $f^{P,i}_z()$, $f^{P,i}_c()$, and $f^{P,i}_{zc}()$ are the corresponding modulation functions, as exemplified in Fig. 4. In general, with a higher PQSM measure, $f^{P,i}_t()$, $f^{P,i}_z()$, and $f^{P,i}_c()$ take lower values, and $f^{P,i}_{zc}()$ takes a higher value.

TABLE V
VALUES OF PARAMETER ϕ_* IN MODULATION FUNCTIONS

functions	ϕ_*
$f_{T_0}^{P,i}()$	0.20
$f_{w_s}^{P,i}()$	0.14
$f_{f_0}^{P,i}()$	0.14
$f_z^{P,i}()$	0.12
$f_c^{P,i}()$	0.12
$f_t^{P,i}()$	0.12
$f_{zc}^{P,i}()$	0.12

For PQSM-modulated JND Model 2, (6) can be rewritten with the (a,b)th DCT block of ith channel as

$$T^{P,i}(u,v,w,a,b) = T_0^{P,i}(a,b)\cdot T_w^{P,i}(w,a,b) \cdot T_f^{P,i}(u,v,a,b)\cdot T_a^i(u,v) \quad (30)$$

where the modulatory aftereffect on contrast orientation is ignored, and

$$T_0^{P,i}(a,b) = T_0^i \cdot f_{T_0}^{P,i}(a,b)) \quad (31)$$

$T_w^{P,i}(w,a,b)$ and $T_f^{P,i}(u,v,a,b)$ can be yielded by substituting

$$w_s^{P,i}(a,b) = w_s^i \cdot f_{w_s}^{P,i}(a,b) \quad (32)$$

$$f_0^{P,i}(a,b) = f_0^i \cdot f_{f_0}^{P,i}(a,b) \quad (33)$$

for w_s and f_0 in (7) and (8). $f_{T_0}^{P,i}()$, $f_{w_s}^{P,i}()$, and $f_{f_0}^{P,i}()$ are the corresponding modulation functions, as exemplified in Fig. 4. For $f_{w_s}^{P,i}()$, $f_{f_0}^{P,i}()$, and $f_{zc}^{P,i}()$, the modulation functions also can be expressed as

$$f^{P,*}(p) = \begin{cases} M_*^{\max}, & p \geq P_{\max} \\ 1+(p-1)\times\phi_*, & P_{\min} < p < P_{\max} \\ M_*^{\min}, & p \leq P_{\min} \end{cases} \quad (34)$$

and for $f_{T_0}^{P,i}()$, $f_z^{P,i}()$, $f_c^{P,i}()$, and $f_t^{P,i}()$, they can be expressed as

$$f^{P,*}(p) = \begin{cases} M_*^{\min}: & p \geq P_{\max} \\ 1-(p-1)\times\phi_*: & P_{\min} < p < P_{\max} \\ M_*^{\max}: & p \leq P_{\min} \end{cases} \quad (35)$$

where ϕ_* is a factor to control the slope for each modulation function.

With higher PQSM measure, the turning point of temporal and spatial contrast masking curves ($w_s^{P,i}$ and $f_0^{P,i}$) are pushed to higher frequencies, and the luminance adaptation tolerances are reduced. Itti's experimental results mentioned in Section II-A.1 have been used in determining the actual maximum and minimum values of each modulation function. The ϕ_* value of each modulation function are obtained by tuning the combined effect to fit Itti's experimental results, as shown in Table V.

V. PQSM-BASED VIDEO QUALITY METRICS (VQMS)

It is expected that the PQSMs modulation on visual sensitivity models enhances the performance of different VQMs. To demonstrate this, two VQMs are used in this paper: VQM A is a PQSM modulated MSE metric, and VQM B is the PQSM modulated Wang's SSIM [16]. Only luminance component has been used for the sake of efficiency, since luminance plays a much more important role in human visual perception than chrominance components.

Let I, $\hat{I}$, and Θ^P denote the original video sequence, the degraded video sequence, and the JND profile estimated by PQSM-modulated JND Model 1, respectively. The perceptual distortion of VQM A can be expressed as

$$E_A = \sum_{(x,y,t)} f_A\left(\left|I(x,y,t)-\hat{I}(x,y,t)\right|, \Theta^P(x,y,t)\right) \quad (36)$$

where

$$f_A(a,b) = \begin{cases} \frac{a}{b}-1, & \text{if } a > b \\ 0, & \text{if } a \leq b. \end{cases} \quad (37)$$

In (36), any distortion below the detectability threshold $\Theta^P(x,y,t)$ is excluded from accumulation for the visual distortion score. Equations (36) and (37) are enhanced formulation to Chou and Li's peak signal-to-perceptual noise ratio (PSPNR) [8], since an above-threshold error is scaled by the threshold.

Because SSIM values are block-based [16], the perceptual distortion of VQM B can be expressed for (r,q)th block as

$$E_B = \sum_{(r,q,t)} f_B\left(\mathrm{SSIM}(r,q,t), \max_{(x',y')\in\Re^{r,q}}\left(\left|I(x',y',t) - \hat{I}(x',y',t)\right|\right), \Theta_b^P(r,q,t), \mathrm{pqsm}_b(r,q,t)\right) \quad (38)$$

where

$$f_B(a,b,c,d) = \begin{cases} a\cdot(d-\beta)^\tau, & \text{if } b > c \\ 0, & \text{if } b \leq c \end{cases} \quad (39)$$

where $\Re^{r,q}$ represents the (r,q)th block, Θ_b^P and pqsm_b denote the block-based Θ^P and pqsm. $\tau = 1.2$ and $\beta = 0.4$ are the constants to map PQSM values into an appropriate range nonlinearly. With the choice of $f_B(a,b,c,d)$, VQM B accounts for the effect of both PQSM and $\Theta_b^P(r,q,t)$: If the maximum block-based error is below $\Theta_b^P(r,q,t)$, it is not accumulated for E_B; otherwise, the PQSM values are nonlinearly and monotonously scaled as the weighting for visual annoyance measurement.

VI. EXPERIMENTAL RESULTS

A. PQSM Estimation

The intermediate and final PQSM estimation results on the 30th frame of *Suzie*, *Autumn leaves* and *Foreman* video test sequences are shown in Figs. 5–7, respectively. In Fig. 5, all detected stimuli and features have contributed in forming the final PQSM indicating the attentional levels in the image, largely in line with the human perception. The shapes of the face and the eyes are somewhat arbitrarily set, since Rowley's algorithm [94] only indicates the upper right and bottom left corners of a face, and the centers of eyes. This simple representation of face and eyes is not accurate, but sufficient for the application. Fig. 6 shows the results for a natural scene, where the detected stimuli and features except for human faces contribute to the final result.

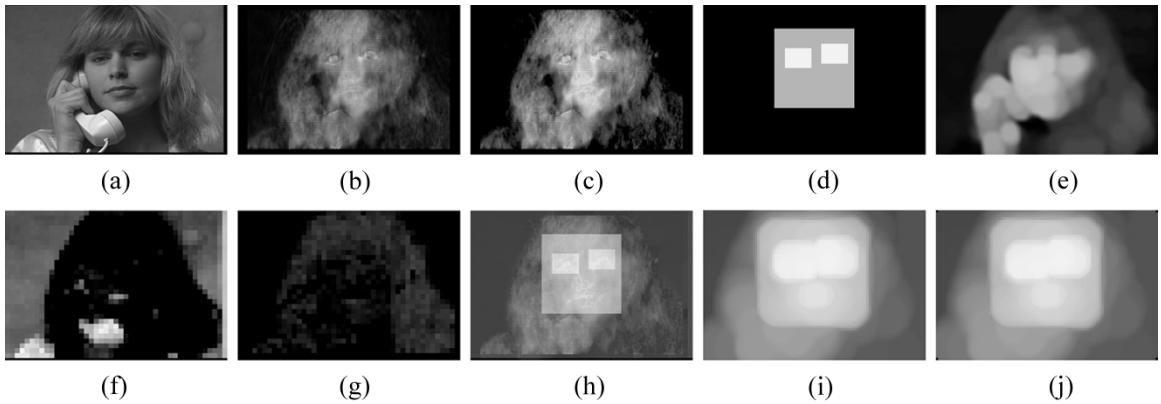


Fig. 5. Example 1 on PQSM estimation: (a) 30th frame of video sequence "Suzie"; (b) absolute motion map; (c) relative motion contrast stimulus; (d) face-eye stimulus; (e) skin color stimulus; (f) color contrast stimulus; (g) texture stimulus; (h) integrated visual attentional map; (i) block-based attentional map after post-processing; and (j) the final PQSM with motion suppression.

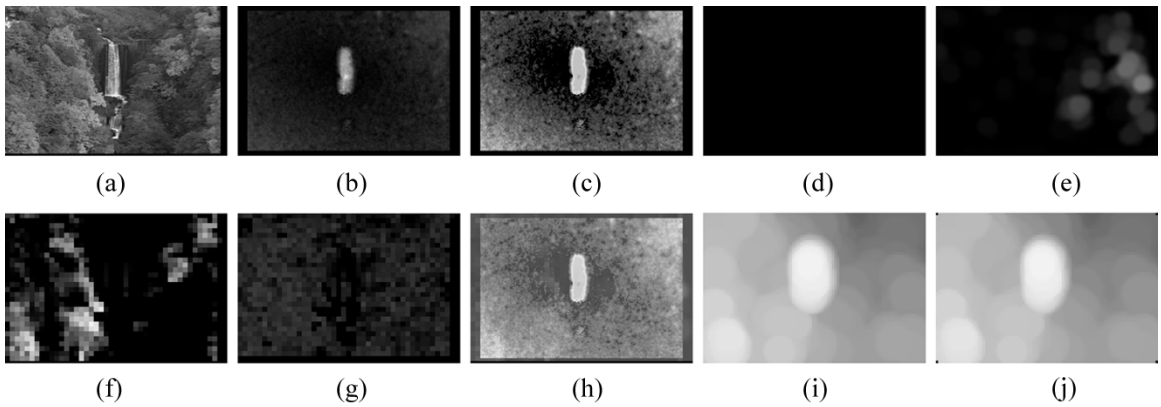


Fig. 6. Example 2 on PQSM estimation: (a) 30th frame of video sequence "Autumn leaves"; (b) absolute motion map; (c) relative motion contrast stimulus; (d) face-eye stimulus; (e) skin color stimulus; (f) color contrast stimulus; (g) texture stimulus; (h) integrated visual attentional map; (i) block-based attentional map after post-processing; and (j) the final PQSM with motion suppression.

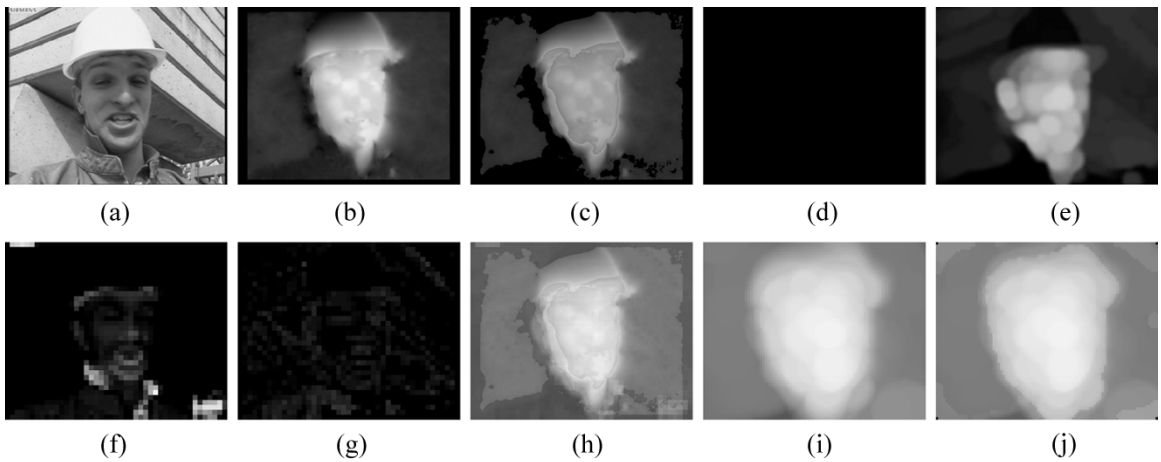


Fig. 7. Example 3 on PQSM estimation: (a) 30th frame of video sequence "Foreman"; (b) absolute motion map; (c) relative motion contrast stimulus; (d) face-eye stimulus; (e) skin color stimulus; (f) color contrast stimulus; (g) texture stimulus; (h) integrated visual attentional map; (i) block-based attentional map after post-processing; and (j) the final PQSM with motion suppression.

It is worth noting that some false skin color has been detected but does not bring about significant impact toward the final PQSM. In Fig. 7, face detection is a failure (face detection is currently still a challenging task because of the variability in scale, location, orientation, and pose [97]), but the skin color and relative motion contrast still lead to reasonable alignment with the HVS observation. Inclusion of mutlistimuli/features helps to increase the application scopes and algorithm robustness. In general, relative motion, face and skin color are the more important factors to the resultant PQSM.

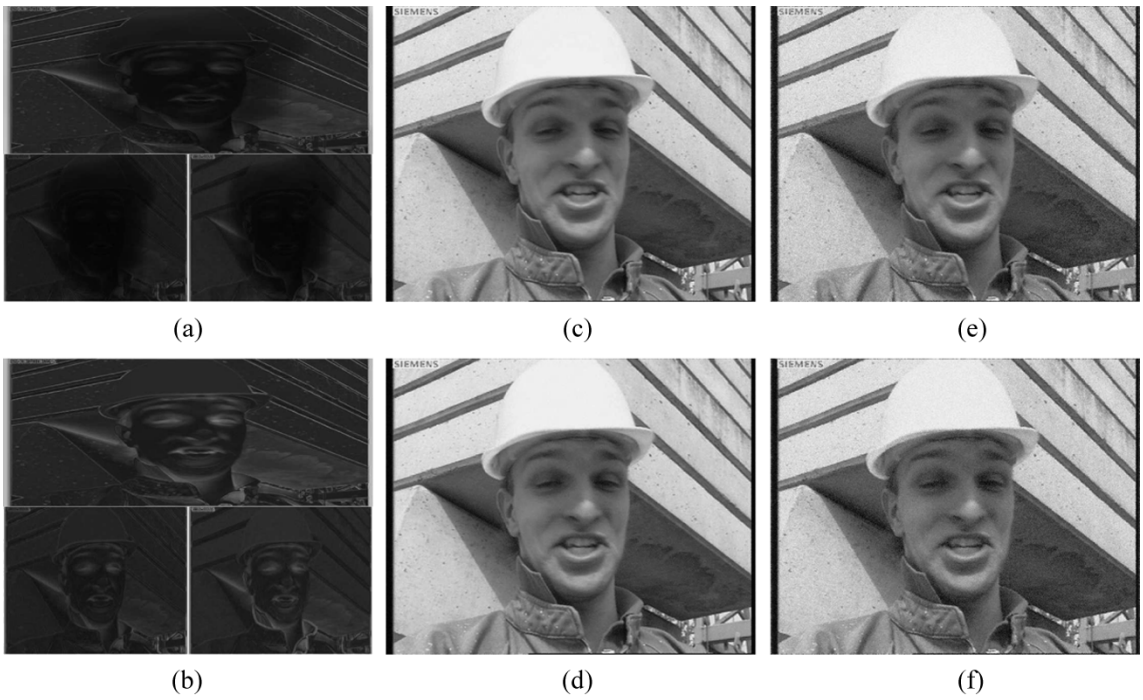


Fig. 8. Embedding noise into the 30th frame of video "Foreman": (a) JND map generated by the proposed PQSM-modulated JND Model 1; (b) JND map generated by Yang's model; (c) noise-injected image based on PQSM-modulated JND Model 1 with $\mathrm{PSNR} = 32.27$ dB; (d) noise-injected image based on Yang's JND model with $\mathrm{PSNR} = 32.84$ dB; (e) noise-injected image based on PQSM-modulated JND Model 1 with $\mathrm{PSNR} = 24.61$ dB; and (f) noise-injected image based on Yang's JND model with $\mathrm{PSNR} = 25.18$ dB.

B. PQSM-Modulated JND Models

To evaluate the performance of proposed PQSM-modulated JND Models 1 and 2 against their original models, noise is injected into video according to the JND models. For a pixel-based JND model (i.e., the original Yang's model [9] or PQSM-modulated JND Model 1), a noise-injected image frame can obtained as

$$\hat{I}^i(x,y,t) = I^i(x,y,t) + \varepsilon_\Theta \cdot \Theta_o^i(x,y,t) \cdot \mathrm{sgn}\left(\mathrm{random}(x,y,t)\right) \quad (40)$$

where Θ_o^i is Θ^i for Yang's model, or $\Theta^{P,i}$ for PQSM-modulated JND Model 1; $\varepsilon_\Theta \leq 1$ for perceptually lossless noise (if the visibility threshold is correctly determined), and $\varepsilon_\Theta > 1$ for perceptually lossy noise; $\mathrm{random}()$ gives a random number, and is used here just to control the sign of the associated term in (40) so that no artificial pattern is added in the spatial space and along the temporal axis.

Such a noise injection scheme can be used to examine the performance of $\Theta^{P,i}(x,y,t)$ against $\Theta^i(x,y,t)$. A more accurate JND model should derive a noise injected image (or video) with better visual quality under the same level of noise (controlled by ε_Θ), because it is capable of shaping more noise onto the less perceptually significant regions in the image. PSNR is used here just to denote the injected noise level under different test conditions. With the same PSNR, the JND model relating to a better subjective visual quality score is a better model. Alternatively, with the same perceptual visual quality score, the JND model relating to a lower PSNR is the better model.

For a DCT-based JND profile $T_o^i(u,v,w)$, the injected noise is obtained as

$$d^i(u,v,w) = \varepsilon_T \cdot T_o^i(u,v,w) \cdot \mathrm{sgn}\left(\mathrm{random}(u,v,w)\right) \quad (41)$$

where T_o^i is $T^i(u,v,w)$ for the original Watson's model [5], or $T^{P,i}(u,v,w)$ for PQSM-modulated JND Model 2; ε_T and $\mathrm{random}()$ have similar meanings as in (40).

Therefore, the degraded video sequence is

$$\hat{I}^i(x,y,t) = \mathrm{DCT}^{-1}\left(\mathrm{DCT}\left(I^i(x,y,t)\right) + d^i(u,v,w)\right) \quad (42)$$

where $\mathrm{DCT}()$ and $\mathrm{DCT}^{-1}()$ denote DCT transform and inverse DCT transform, respectively.

To compare Yang's JND model and the proposed PQSM-modulated JND Model 1, the noise injected images are used for four test sequences: *Suzie*, *Autumn leaves*, *Foreman*, and *Harp*. An example of the generated JND profiles and noise injected images are shown in Fig. 8. Fig. 8(a) shows the generated JND profile by PQSM-modulated JND model 1 while Fig. 8(b) shows the generated JND profile by Yang's JND model, with the upper, lower left, and lower right parts being the Y, Cb, and Cr components, respectively; Fig. 8(c) and (e) are the noise-injected images by PQSM-modulated JND Model 1 with different PSNR levels; Fig. 8(d) and (f) are the noise-injected images by Yang's Model, with similar or slightly higher PSNRs in comparison with Fig. 8(c) and (e). Fig. 9 gives a close-up view in the most sensitive region (i.e., with highest PQSM values) of Fig. 8 for better visualization, and, as can be seen, PQSM-modulated JND Model 1 yields better visual quality in the noise-injected images.

PQSM-modulated JND Model 2 is compared with Watson's JND model in the similar manner. Fig. 10 shows a close-up view

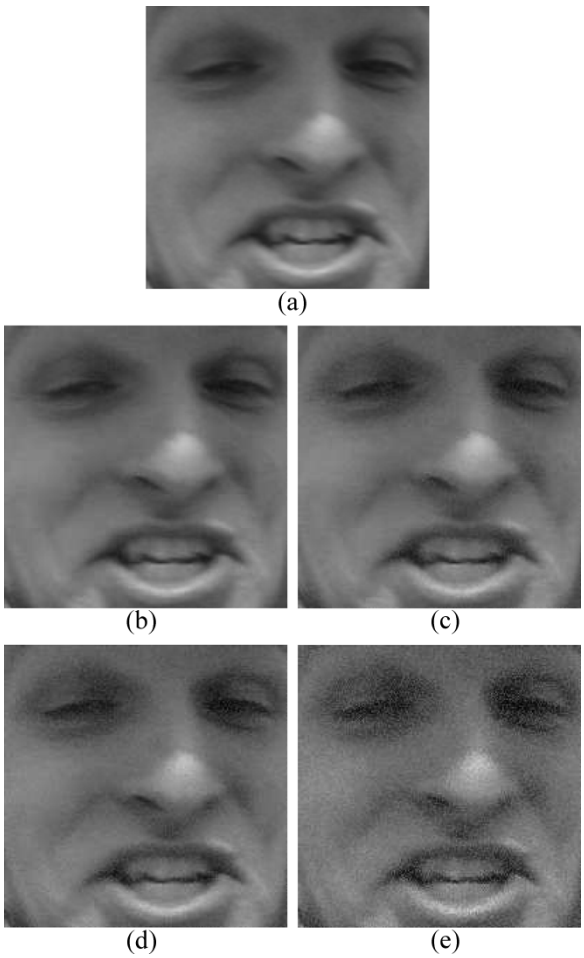


Fig. 9. Details of the most sensitive (with the highest PQSM values) region of Fig. 7: (a) original image; (b) noise-injected image based on PQSM-modulated JND Model 1 with $\mathrm{PSNR} = 32.27$ dB; (c) noise-injected image based on Yang's JND model with $\mathrm{PSNR} = 32.84$ dB; (d) noise-injected image based on PQSM-modulated JND Model 1 with $\mathrm{PSNR} = 24.61$ dB; and (e) noise-injected image based on Yang's JND model with $\mathrm{PSNR} = 25.18$ dB.

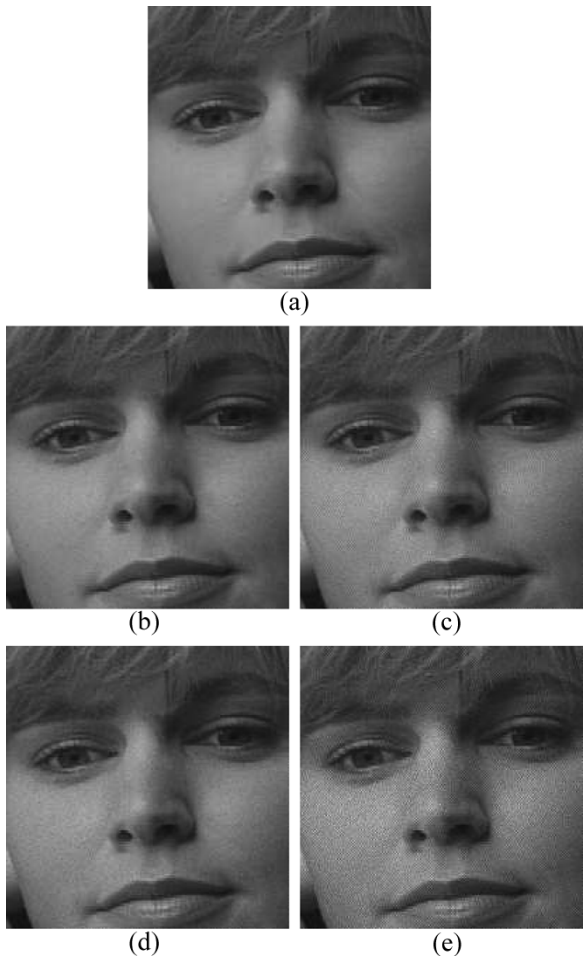


Fig. 10. Details of most sensitive (with the highest PQSM values) region of sequence "Suzie": (a) original image; (b) noise-injected image based on PQSM-modulated JND Model 2 with $\mathrm{PSNR} = 31.26$ dB; (c) noise-injected image based on Watson's JND model with $\mathrm{PSNR} = 31.37$ dB; (d) noise-injected image based on PQSM-modulated JND Model 2 with $\mathrm{PSNR} = 27.53$ dB; and (e) noise-injected image based on Watson's JND model with $\mathrm{PSNR} = 27.59$ dB.

in the most sensitive region of the 30th frame of *Suzie* sequence under two PSNR conditions. As expected, PQSM-modulated JND Model 2 yields better visual quality in the noise-injected images.

To confirm the above-mentioned visual quality observation, formal subjective viewing tests have been conducted for the noise-injected sequences based on the two pairs of JND models (namely, Yang's JND model and PQSM-modulated JND Model 1, Watson's JND model and PQSM-modulated JND Model 2), with the noise conditions listed in the upper portions of Tables VI and VII. Each display and scoring session for a pair of noise-injected sequences generated by a pair of JND models is organized as: *Video Sequence I, two seconds of grey screen, Video Sequence II, two seconds of grey screen, Video Sequence I, two seconds of grey screen, Video Sequence II, two seconds of grey screen, scoring*. Both the display order of the sequences in a session, the order of noise-injection conditions, and the order of th four test sequences were randomized for subjects. The preference opinion scores (POSs) are given by subjects to indicate the quality comparison: 1) $Q_I > Q_{II}$, the visual quality of Sequence I is better than Sequence II; 2) $Q_I = Q_{II}$, viewer can not decide which sequence has better quality; and 3) $Q_I < Q_{II}$, the visual quality of Sequence II is better than Sequence I.

Eight subjects (four of them are with average image processing knowledge and the rest are naive) were involved in the experiments. Their eyesight is either normal or has been corrected to be normal with spectacles. The subjective visual quality assessment was performed in a typical laboratory environment with normal fluorescent ceiling light, using a $21''$ EIZO T965 professional color monitor with resolution of 1600×1200, screen refresh rate at 85 Hz. The luminance, Gamma curve and Saturation setup are using the "Movie" display mode. The viewing distance is approximately four times of the image height.

Table VI is the viewing results between Yang's JND model and PQSM-modulated JND Model 1, and Table VII is the results between Watson's JND model and PQSM-modulated JND Model 2. In Table VI, $Q_{\mathrm{Yang}} > Q_{\mathrm{PQSM1}}$ indicates that the quality of the sequence with Yang's JND Model is better, $Q_{\mathrm{Yang}} = Q_{\mathrm{PQSM1}}$ indicates that two sequences has the same quality, and $Q_{\mathrm{Yang}} < Q_{\mathrm{PQSM1}}$ indicates that the quality of the sequence with PQSM-modulated JND Model 1 is better.

TABLE VI
POSs OF THE NOISE-INJECTED IMAGES GENERATED BY YANG'S JND MODEL AND PQSM-MODULATED JND MODEL 1

	'Foreman'		'Harp'		'Autumn leaves'		'Suzie'	
$PSNR_{Yang}$	$32.84dB$	$25.18dB$	$32.93dB$	$25.38dB$	$32.11dB$	$27.71dB$	$32.84dB$	$25.18dB$
$PSNR_{PQSM1}$	$32.27dB$	$24.61dB$	$32.90dB$	$25.06dB$	$32.11dB$	$27.64dB$	$32.27dB$	$24.61dB$
$Q_{Yang} > Q_{PQSM1}$	0	0	0	1	0	0	0	0
$Q_{Yang} = Q_{PQSM1}$	0	0	2	1	4	2	0	1
$Q_{Yang} < Q_{PQSM1}$	8	8	6	6	4	6	8	7

TABLE VII
POSs OF THE NOISE-INJECTED IMAGES GENERATED BY WATSON'S JND MODEL AND PQSM-MODULATED JND MODEL 2

	'Foreman'		'Harp'		'Autumn leaves'		'Suzie'	
$PSNR_{Watson}$	$30.48dB$	$27.56dB$	$33.59dB$	$27.56dB$	$32.66dB$	$27.54dB$	$31.37dB$	$27.59dB$
$PSNR_{PQSM2}$	$30.20dB$	$27.50dB$	$33.43dB$	$27.40dB$	$32.45dB$	$27.40dB$	$31.26dB$	$27.53dB$
$Q_{Watson} > Q_{PQSM2}$	0	0	1	2	1	1	0	0
$Q_{Watson} = Q_{PQSM2}$	0	1	2	3	3	1	1	2
$Q_{Watson} < Q_{PQSM2}$	8	7	5	3	4	6	7	6

TABLE VIII
PERFORMANCE COMPARISON FOR VQM A AND VQM B WITH VQEG PHASE-I DATA SET

	PSNR			VQM A			SSIM			VQM B		
	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
50Hz	0.786	0.810	0.728	0.820	0.812	0.567	-	-	-	0.889	0.859	0.539
60Hz	0.760	0.711	0.583	0.795	0.762	0.628	-	-	-	0.914	0.897	0.556
All	0.779	0.786	0.678	0.812	0.805	0.603	0.849	0.812	0.578	0.895	0.871	0.541

Similar notations are used in Table VII. Subjective viewing results confirm that accounting for the modulatory aftereffects of visual attention and motion suppression enhances the performance of JND models.

C. PQSM-Based VQMs

In Section V, VQM A and VQM B have been formulated after the consideration of the modulatory aftereffects of visual attention and motion suppression. In this section, their performance is compared against their original forms before such modulatory aftereffects are included, that is, the PSNR (equivalent to MSE) measure and the SSIM measure [16]. The performance has been evaluated, using the most extensive publicly-accessible database for visual quality assessment, VQEG Phase-I test set [98], which includes 20 SDTV test sequences, their 320 decoded sequences and the associated subjective rating results [difference mean opinion scores (DMOSs)].

In [98], a three-parameter logistic function is used to estimate the predicted DMOS_p from VQR (video quality rating) and the output of a VQM

$$\text{DMOS}_p = \frac{b1}{1 + \exp\left(-b2 * (\text{VQR} - b3)\right)} \tag{43}$$

where $b1$, $b2$, and $b3$ are parameters derived from fitting VQR to DMOS. Three methods are adopted in VQEG [98] to evaluate the prediction accuracy of VQMs.

- Method 1 (M_1): Pearson linear correlation coefficient between DMOS_p and DMOS.
- Method 2 (M_2): Spearman rank order correlation coefficient between DMOS_p and DMOS.
- Method 3 (M_3): Outlier ratio of *outlier sequences* to the total number of testing sequences.

The higher M_1 and M_2 are and the lower M_3 is, the better match with DMOS a VQM achieves. In the ideal match, M_1 and M_2 have the value of 1 and M_3 is 0.

The experimental results are listed in Table VIII, comparing VQM A with the PSNR measure and VQM B with the SSIM measure. The three major test groups of the VQEG Phase-I data have been used: 1) the 50-Hz set with 180 decoded sequences in PAL format; 2) the 60-Hz set with 180 decoded sequences in NSTC format; and 3) all 320 decoded sequences. The PSNR results are obtained from [98], while the SSIM results are obtained from [16]. As can be seen, VQM A outperforms the PSNR in all three test groups and all three evaluation methods (M_1 to M_3). For the comparison between VQM B and the SSIM, although the SSIM results of the 50- and 60-Hz data sets are unavailable, we can still see that the accuracy of VQM B is improved over the SSIM assessment for all 320 decoded sequences. As for the comparison between VQM A and VQM B, the performance of VQM B is better than that of VQM A, because VQM B accounts for structural errors and visual annoyance in decoded images.

VII. CONCLUSION AND FUTURE WORK

This paper first gives a comprehensive review on the existing research effort related to the studies of visual attention, computational models of visual sensitivity, and perceptual quality/distortion metrics. We then demonstrate that accounting for visual attention's modulatory aftereffects improves the visual sensitivity and the quality/distortion evaluation. In this paper, we build the model according to the basic ideas of visual attention

that are just sufficient to the problems being tackled and also include important cognitive features for more effective applications in practice.

A numerical measure of visual attention's modulatory after-effects in an image or video, PQSM, is proposed. The devised PQSM estimation algorithm detects motion, color contrast and texture contrast as *bottom-up* stimuli, and also skin color and face *top-down* features, in this paper. Two PQSM-modulated visual sensitivity (JND) models and two PQSM-modulated VQMs are also presented for evaluating the proposed PQSM. Extensive experimental results with subjective viewing confirm the improvement on both JND determination and visual quality gauge.

The proposed PQSM basically provides local perceptual cues of significance toward visual quality. How picture quality is gauged shapes the design, implementation and optimization of most visual processing tasks. Therefore, apart from the demonstrated applications in JND determination and visual quality gauge, the proposed PQSM can facilitate perceptually-optimized image/video compression, watermarking, error resilience and many other processes.

As for the possible further work, more features/stimuli may be included, especially semantic features and auditory stimuli (auditory stimuli play an important role on spatial attention selection [99]). Temporal effect of visual attention is another direction for improvement [96]. It is believed that *bottom-up* stimuli are dominative on visual attention during a short period (100∼800 ms) when scene appears/changes, and afterwards, *top-down* features will possibly dominate. On the other hand, after a period (several seconds to a few minutes, depending on the type of stimulus), a stimulus' attentional level decreases with the time of its appearance in the visual field.

ACKNOWLEDGMENT

The authors would like to thank Prof. T. Kanade, Dr. T. Moriyama, and Dr. H. A. Rowley of Carnegie Mellon University for providing their face detection codes, and Assoc. Prof. M. J. Black of Brown University for providing source code of his optical flow algorithm.

REFERENCES

[1] S. Daly, *The Visible Differences Predictor: An Algorithm for The Assessment of Image Fidelity*. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[2] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for blocking artifacts in image based on human visual sensitivity," *IEEE Trans. Image Process.*, vol. 4, no. 6, pp. 713–724, Jun. 1995.

[3] J. Malo, J. Gutiérrez, I. Epifanio, F. J. Ferri, and J. M. Artigas, "Perceptual feedback in multigrid motion estimation using an improved Dct quantization," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1411–1427, Oct. 2001.

[4] S. Winkler, "Vision Models and Quality metrics for image processing applications," Ph.D. dissertation, Signal Process. Lab., Ecole Polytechnique Federale De Lausanne (EPFL), Swiss Fed. Inst. Technol., Lausanne, Switzerland, Dec. 2000.

[5] A. B. Watson, J. Hu, and J. F. McGowan III, "DVQ: A digital video quality metric based on human vision," *J. Elect. Imag.*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[6] H. Yee, S. Pattanaik, and D. P. Greenberg, "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments," *ACM Trans. Graphics*, pp. 39–65, 2001.

[7] S.-S. Kuo and J. D. Johnston, "Spatial noise shaping based on human visual sensitivity and its application to image coding," *IEEE Trans. Image Process.*, vol. 11, no. 5, pp. 509–517, May 2002.

[8] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 5, pp. 467–476, May 1995.

[9] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong, and S. S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images," in *Proc. ICASSP*, vol. 3, Apr. 2003, pp. 609–612.

[10] A. J. Ahumada and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," in *Proc. SPIE Human Vision, Visual Processing, and Digital Display III*, vol. 1666, 1992, pp. 365–374.

[11] P. T. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. Int. Conf. Image Processing*, vol. 2, Nov. 1994, pp. 982–986.

[12] J. Lubin, M. H. Brill, and A. P. Pica, "Method and apparatus for estimation digital video quality without using a reference video," U.S. Patent 6 285 797, Sep. 2001.

[13] P. le Callet and D. Barba, "Perceptual color image quality metric using adequate error pooling for coding scheme evaluation," in *Proc. SPIE Human Vision and Electronic Imaging VII*, vol. 4662, B. Rogowitz and T. N. Pappas, Eds., San Jose, CA, Jun. 2002, pp. 173–180.

[14] J. Li, G. Chen, Z. Chi, and C. Lu, "Image coding quality assessment using fuzzy integrals with a three-component image model," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 1, pp. 99–106, Jan. 2004.

[15] H. R. Wu, "A new distortion measure for video coding blocking artifacts," in *Proc. ICCT*, vol. 2, May 1996, pp. 658–661.

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

[17] E. P. Ong, W. S. Lin, Z. K. Lu, S. S. Yao, X. K. Yang, and F. Moschetti, "Low bit rate video quality assessment based on perceptual characteristics," in *Proc. Int. Conf. Image Processing*, vol. 3, Sep. 2003, pp. 189–192.

[18] M. Yuen and H. R. Wu, "A survey of Hbrid MC/DPCM/DCT video coding distortions," *Signal Process.*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[19] M. M. Chun and J. M. Wolfe, "Visual attention," in *Blackwell Handbook of Perception*, B. Goldstein, Ed. Oxford, UK: Blackwell, 2001, pp. 272–310.

[20] W. James, *The Principles of Psychology*. Cambridge, MA: Harvard Univ. Press, 1890.

[21] J. K. Tsotsos, "Motion understanding: task-directed attention and representations that like perception with action," *Int. J. Comput. Vis.*, vol. 45, no. 3, pp. 265–280, Dec. 2001.

[22] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, Dept. Computat. Neur. Syst., California Inst. Technol., Pasadena, Jan. 2000.

[23] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neurosci.*, vol. 3, no. 3, pp. 284–291, Mar. 2000.

[24] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu. Rev. Neurosci.*, vol. 23, pp. 315–341, Mar. 2000.

[25] R. D. Wright and C. M. Richard, "Sensory mediation of stimulus-driven attentional capture in multiple-cue displays," *Perception Psychophys.*, vol. 65, no. 6, pp. 925–938, 2003.

[26] D. M. Lloyd, N. Merat, F. McGlone, and C. Spence, "Crossmodal links between audition and touch in covert endogenous spatial attention," *Perception Psychophys.*, vol. 65, no. 6, pp. 901–924, 2003.

[27] S. J. Luck and M. A. Ford, "On the role of selective attention in visual perception," *Proc. Nat. Acad. Sci.*, vol. 95, no. 3, pp. 825–830, Feb. 1998.

[28] L. Itti, J. Braun, and C. Koch, "Modeling the modulatory effect of attention on human spatial vision," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, vol. 14.

[29] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[30] W. Osberger, "Perceptual vision models for picture quality assessment and compression applications," Ph.D. dissertation, Space Centre for Satellite Navigation, Queensland Univ. Technol., Brisbane, Australia, Mar. 1999.

[31] O. L. Meur, P. Le Callet, D. Barba, D. Thoreau, and E. Francois, "From low level perception to high level perception, a coherent approach for visual attention modeling," in *Proc. SPIE Human Vision and Electronic Imaging IX*, vol. 5292, B. Rogowitz and T. N. Pappas, Eds., San Jose, CA, Jan. 2004, pp. 284–295.

[32] B. J. Murphy, "Pattern thresholds for moving and stationary gratings during smooth eye movement," *Vis. Res.*, vol. 18, no. 5, pp. 521–530, 1978.

[33] L. Michels and M. Lappe, "Contrast dependency of saccadic compression and suppression," *Vis. Res.*, vol. 44, no. 20, pp. 2327–2336, Sep. 2004.

[34] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision Models and Applications to Image and Video Processing*, C. J. van den Branden Lambrecht, Ed. Boston, MA: Kluwer, 2001, ch. 9, pp. 179–200.

[35] M. Eimer, "An ERP study of sustained spatial attention to stimulus eccentricity," *Biol. Psych.*, vol. 52, no. 3, pp. 205–220, Mar. 2000.

[36] Y.-J. Luo, P. M. Greenwood, and R. Parasuraman, "Dynamics of the spatial scale of visual attention revealed by brain event-related potentials," *Cogn. Brain Res.*, vol. 12, no. 3, pp. 371–381, Dec. 2002.

[37] H. A. Allen and T. Ledgeway, "Attentional modulation of threshold sensitivity to first-order motion and second-order motion patterns," *Vis. Res.*, vol. 43, no. 27, pp. 2927–2936, Dec. 2003.

[38] H. D. Lange, "Research into the dynamic nature of the human fovea cortex system with intermittent and modulated light. I. Attenuation characteristics with white and colored light," *J. Opt. Soc. Amer.*, vol. 48, pp. 777–784, 1958.

[39] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Bellingham, WA: SPIE, 1999.

[40] M. J. Nadenau, J. Reichel, and M. Kunt, "Wavelet-based color image compression: exploiting the contrast sensitivity function," *IEEE Trans. Image Process*, vol. 12, no. 1, pp. 58–70, Jan. 2003.

[41] R. L. D. Valois and K. K. D. Valois, *Spatial Vision*. Oxford, U.K.: Oxford Univ. Press, 1988.

[42] A. 'P. Ginsburg, "Vision channels, contrast sesitivity, and functional vision," in *Proc. SPIE Human Vision and Electronic Imaging IX*, vol. 5292, B. Rogowitz and T. N. Pappas, Eds., San Jose, CA, Jan. 2004, pp. 15–25.

[43] P. Le Callet, A. Saadane, and D. Barba, "Interactions of chromatic components on the perceptual quantization of the achromatic component," in *Proc. SPIE Human Vision and Electronic Imaging*, vol. 3644, B. Rogowitz and T. N. Pappas, Eds., San Jose, CA, May 1999, pp. 121–128.

[44] P. Le Callet and D. Barba, "Robust approach for color image quality assessment," in *Proc. Int. Conf. Visual Communications and Image Processing*, vol. 5150, Jul. 2003, pp. 1573–1581.

[45] H. S. Bashinski and V. R. Bacharach, "Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations," *Perception Psychopys.*, vol. 28, no. 3, pp. 241–280, 1980.

[46] H. Muller and J. M. Findlay, "Sensitivity and criterion effects in the spatial cueing of visual attention," *Perception Psychopys.*, vol. 42, no. 4, pp. 383–399, 1987.

[47] V. M. Ciaramitaro, E. L. Cameron, and P. W. Glimcher, "Stimulus probability directs spatial attention: an enhancement of sensitivity in humans and monkeys," *Vis. Res.*, vol. 41, no. 1, pp. 57–75, Jan. 2001.

[48] T. D. Tran and R. Safranek, "A locally adaptive perceptual masking threshold model for image coding," in *Proc. ICASSP*, vol. 4, 1996, pp. 1883–1886.

[49] H. Y. Tong and A. N. Venetsanopoulos, "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking," in *Proc. Int. Conf. Image Process.*, vol. 3, 1998, pp. 428–432.

[50] C. H. Chou and C. W. Chen, "A perceptually optimized 3-D subband image codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 2, pp. 143–156, Feb. 1996.

[51] M. S. Moore, J. Foley, and S. K. Mitra, "Detectability and annoyance value of MPEG-2 artifacts inserted in uncompressed video sequences," in *Proc. SPIE Human Vision and Electronic Imaging V*, vol. 3959, B. E. Rogowitz and T. N. Pappas, Eds., San Jose, CA, Jan. 2000, pp. 99–110.

[52] S. Treue and J. C. M. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, pp. 575–579, Jun. 1999.

[53] T. A. W. Visser and J. T. Enns, "The role of attention in temporal integration," *Perception*, vol. 30, no. 2, pp. 135–145, Feb. 2001.

[54] L. Paul and P. G. Schyns, "Attention enhances feature integration," *Vis. Res.*, vol. 43, no. 17, pp. 1793–1798, Aug. 2003.

[55] J. Saiki, "Spatiotemporal characteristics of dynamic feature binding in visual working memory," *Vis. Res.*, vol. 43, no. 20, pp. 2107–2123, Sep. 2003.

[56] P. Verghese and D. G. Pelli, "The information capacity of visual attention," *Vis. Res.*, vol. 32, no. 5, pp. 983–995, May 1992.

[57] J. A. Solomon, "The effect of spatial cues on visual sensitivity," *Vis. Res.*, to be published.

[58] R. VanRullen and T. Dong, "Attention and scintillation," *Vis. Res.*, vol. 43, no. 21, pp. 2191–2196, Sep. 2003.

[59] C. J. Downing and S. Pinker, "The spatial structure of visual attention," in *Attention and Performance XI*, M. I. Posner and O. S. Marin, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1985, pp. 171–187.

[60] C. W. Eriksen and J. D. S. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception Psychophys.*, vol. 40, pp. 225–240, 1986.

[61] N. F. Nielsen, J. Michelsen, J. A. Michelsen, and T. Schneider, "Numerical calculation of electrostatic field surrounding a human head in visual display environments," *J. Elect.*, vol. 36, no. 3, pp. 209–223, Jan. 1996.

[62] V. James, J. Kearsley, T. Irving, Y. Amemiya, and D. Cookson, "Change-blindness as a result of 'mudsplashes'," *Nature*, vol. 398, p. 34, Mar. 1999.

[63] A. Sahraie, M. Milders, and M. Niedeggen, "Attention induced motion blindness," *Vis. Res.*, vol. 41, no. 13, pp. 1613–1617, Jun. 2001.

[64] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature*, vol. 390, pp. 279–281, 1997.

[65] R. Landman, H. Spekreijsea, and V. A. F. Lamme, "Large capacity storage of integrated objects before change blindness," *Vis. Res.*, vol. 43, no. 2, pp. 149–164, Jan. 2003.

[66] D. Soto and M. J. Blanco, "Spatial attention and object-based attention: A comparison within a single task," *Vis. Res.*, vol. 44, no. 1, pp. 69–81, Jan. 2004.

[67] H.-C. Nothdurft, "Salience from feature contrast: additivity across dimensions," *Vis. Res.*, vol. 40, no. 10–12, pp. 1183–1201, Jun. 2000.

[68] ——, "Salience from feature contrast: temporal properties of saliency mechanisms," *Vis. Res.*, vol. 40, no. 18, pp. 2421–2435, Aug. 2000.

[69] ——, "Salience from feature contrast: variations with texture density," *Vis. Res.*, vol. 40, no. 23, pp. 3181–3200, Jan. 2000.

[70] M. C. Morrone, V. Denti, and D. Spinelli, "Color and luminance contrasts attract independent attention," *Curr. Biol.*, vol. 12, no. 13, pp. 1134–1137, Jul. 2002.

[71] D. E. Broadbent, *Perception and Communication*. London, U.K.: Pergamon, 1958.

[72] C. Koch and S. Ullman, "Shifts in selective visual attention: toward the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.

[73] J. K. Tsotsos, S. M. Culhanea, Y. K. W. Winkya, L. Yuzhonga, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1–2, pp. 507–545, Oct. 1995.

[74] L. Itti, "Visual attention," in *The Handbook of Brain Theory and Neural Networks*, 2nd ed, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 2003, pp. 1196–1201.

[75] T. Koshizen, K. Akatsuka, and H. Tsujino, "A computational model of attentive visual system induced by cortical neural network," *Neurocomput.*, vol. 44–46, pp. 881–887, Jun. 2002.

[76] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cogn. Psych.*, vol. 12, no. 1, pp. 97–136, 1980.

[77] *JPEG 2000 Part I Final Committee Draft Version 1.0*, ISO/IEC JTC 1/SC 29/WG 1 (ITU-T SG8) Std., Mar. 2000.

[78] *JPEG 2000 Part II Final Committee Draft Version 1.0*, ISO/IEC JTC 1/SC 29/WG 1 (ITU-T SG8) Std., Dec. 2000.

[79] C. Christopoulos, A. N. Skodras, and T. Ebrahimi, "JPEG2000 still image coding system: An overview," *IEEE Trans. Consum. Electron.*, vol. 46, no. 4, pp. 1103–1127, Nov. 2000.

[80] M. Reddy, "Perceptually modulated level of detail for virtual environments," Ph.D. dissertation, Dept. Comp. Sci., Univ. Edinburgh, Edinburgh, U.K., 1997.

[81] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

[82] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[83] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 149–162, Feb. 2003.

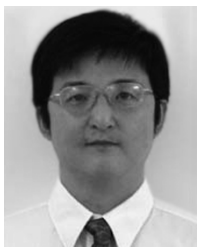
[84] Z. Wang, S. Banerjee, B. L. Evans, and A. C. Bovik, "Generalized bitplane-by-bitplane shift method for JPEG2000 ROI coding," in *Proc. Int. Conf. Image Process.*, vol. 3, Sep. 2002, pp. 81–84.

[85] K. Cater, A. Chalmers, and G. Ward, "detail to attention: exploiting visual tasks for selective rendering," in *Proc. 13th Eurographics Workshop on Rendering*, P. Christensen and D. Cohen-Or, Eds., Jun. 2003, pp. 270–280.

[86] N. Dhavale and L. Itti, "Saliency-based multi-foveated MPEG compression," presented at the IEEE 7th Int. Symp. Signal Processing and Its Applications, Jul. 2003.

[87] X. K. Yang, W. S. Lin, Z. K. Lu, X. Lin, S. Rahardja, E. P. Ong, and S. S. Yao, "Local visual perceptual clues and its use in videophone rate control," presented at the ISCAS, 2004.

[88] Z. K. Lu, W. S. Lin, E. P. Ong, S. S. Yao, and X. K. Yang, "Perceptual-quality significance map (PQSM) and its application on video quality distortion metrics," in *Proc. ICASSP*, vol. 3, Hong Kong, Apr. 2003, pp. 617–620.
[89] Z. K. Lu, W. S. Lin, X. K. Yang, E. P. Ong, and S. S. Yao, "PQSM based RF and NR video quality metrics," in *Proc. SPIE VCIP*, vol. 5150, Jul. 2003, pp. 633–640.
[90] T. Horowitz and A. Treisman, "Attention and apparent motion," *Spatial Vis.*, vol. 8, no. 2, pp. 193–219, 1994.
[91] D. Alais and R. Blake, "Neural strength of visual attention gauged by motion adaptation," *Nature Neurosci.*, vol. 2, no. 11, pp. 1015–1018, Nov. 1999.
[92] M. J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," *Comput. Vis. Image Understand.*, vol. 63, no. 1, pp. 75–104, Jan. 1996.
[93] K. Zhang and J. Kittler, "Global motion estimation and robust regression for video coding," presented at the ICASSP, 1998.
[94] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.
[95] E. R. Simon-Thomas, K. Brodsky, C. Willing, R. Sinha, and R. T. Knight, "Distributed neural activity during object, spatial and integrated processing in humans," *Cogn. Brain Res.*, vol. 16, no. 3, pp. 457–467, May 2003.
[96] G. Deco, O. Pollatos, and J. Zihl, "The time course of selective visual attention: Theory and experiments," *Vis. Res.*, vol. 42, no. 27, pp. 2925–2945, Dec. 2002.
[97] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.
[98] VQEG (Video Quality Expert Group). (2000, Mar.) Final report from the video quality expert group on the validation of objective models of video quality assessment. [Online]. Available: http://www.vqeg.org
[99] M. C. Doyle and R. J. Snowden, "Identification of visual stimuli is improved by accompanying auditory stimuli: the role of eye movements and sound location," *Perception*, vol. 30, no. 7, pp. 795–810, Jul. 2001.



Zhongkang Lu (S'97–M'99–SM'04) received the B.Eng. degree in biomedical engineering from Southeast University, Nanjing, China, in 1993, and the M.Eng. and Ph.D. degrees in electrical engineering from Shanghai Jiaotong University, Shanghai, China, in 1996 and 1999, respectively.

Between 1996 and 1998, he was an exchange student with the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University. From 1999 to 2001, he was a Research Fellow with School of Electrical and Electronic Engineering, Nanyang Technological University. Thereafter, he joined the Institute for Infocomm Research, Singapore, as a Research Scientist. His research interests include computational aspects of human vision, perceptual visual signal processing, pattern recognition, and computer vision.

Weisi Lin (M'92–SM'98) received the B.Sc. and M.Sc. degrees from Zhongshan University, China, in 1982 and 1985, respectively, and the Ph.D. degree from King's College London, London University, London, U.K., in 1992.

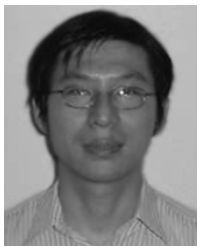
He taught and/or researched at Zhongshan University, China (1982 to 1986), Shantou University, China (1986 to 1988), King's College, London (1990 to 1992), Bath University, Bath, U.K. (1992 to 1993), the National University of Singapore (1993 to 1996), and the Institute of Microelectronics, Singapore (1996 to 2000). He has been the Project Leader of nine successfully delivered industrial-funded projects in the development of digital multimedia-related technologies since 1997. He has published over 70 refereed papers.

Dr. Lin is a Chartered Engineer. He is currently an Associate Lead Scientist and the Project Leader for visual processing at the Institute for Infocomm Research, Singapore. His current research interests include image processing, perceptual visual distortion metrics, perceptual video compression, and multimedia signal processing.

Xiaokang Yang (M'00–SM'04) received the B.Sci. degree from Xiamen University, Xiamen, China, in 1994, the M.Eng. degree from the Chinese Academy of Sciences, Shanghai, in 1997, and the Ph.D. degree from Shanghai Jiaotong University, Shanghai, in 2000.

From September 2000 to March 2002, he was a Research Fellow with the Centre for Signal Processing, Nanyang Technological Institute, Singapore. From April 2002 to October 2004, he was a Research Scientist with the Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the Department of Electronic Engineering, Institute of Image Communication and Information Processing, Shanghai Jiaotong University. He has published over 60 refereed papers and holds six patents. His current research interests include scalable video coding, perceptual video processing, video transmission over networks, and digital television.

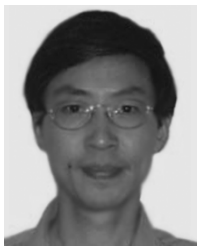
Dr. Yang is a member of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He has received awards from the A-STAR and Tan Kah Kee foundations, as well as the Best Young Investigator Paper Award at the IS&T/SPIE International Conference on Video Communication and Image Processing (VCIP 2003) in perceptual video processing.



EePing Ong received the B.Eng and Ph.D. degrees in electronics and electrical engineering from The University of Birmingham, Birmingham, U.K., in 1993 and 1997, respectively.

From 1997 to 2001, he was with the Institute of Microelectronics, Singapore. Thereafter, he joined the Centre for Signal Processing, Nanyang Technological University, Singapore. Since 2002, he has been with the Institute for Infocomm Research, Singapore, where he is currently a Scientist. His research interests include optical flow, motion estimation, video object segmentation and tracking, perceptual image/video quality metrics and coding.

Dr. Ong is currently Chairman of the IEEE Consumer Electronics Chapter, Singapore.



Susu Yao received the B.Eng. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, and the Ph.D. degree from the National University of Defense Technology, China, in 1983 and 1993, respectively.

He was a Visiting Scholar at Heriot–Watt University, U.K., from 1991 to 1993. From 1993 to 1995, he was a Postdoctoral Fellow and Associate Professor at Southeast University, Nanjing. Since 1996, he has been an Associate Professor and Full Professor at the Nanjing Institute for Communication Engineering. In 2000, he joined the Centre for Signal Processing in Nanyang Technological University, Singapore. His main areas of research interest are image and video compression, wavelet transform, soft computing, image and video post-processing, and perceptual image quality metrics, about which he has published more than 40 papers. He is currently an Associate Lead Scientist at the Institute for Infocomm Research, Singapore.