Rate Control for Videophone Using Local Perceptual Cues

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

***Abstract*—We present a method for extracting local visual perceptual cues and its application for rate control of videophone, in order to ensure the scarce bits to be assigned for maximum perceptual coding quality. The optimum quantization step is determined with the rate-distortion model considering the local perceptual cues in the visual signal. For extraction of the perceptual cues, luminance adaptation and texture masking are used as the stimulus-driven factors, while skin color serves as the cognition-driven factor in the current implementation. Both objective and subjective quality evaluations are given by evaluating the proposed perceptual rate control (PRC) scheme in the H.263 platform, and the evaluations show that the proposed PRC scheme achieves significant quality improvement in block-based coding for bandwidth-hungry applications.**

***Index Terms*—Human visual system (HVS), perception, rate control, video coding, videophone.**

I. Introduction

With the wireless network explosion, low bit-rate video traffic on handheld devices, especially videophone, is likely to represent a significant portion of the data transmitted via the wireless network in the near future. A typical videophone scene is basically composed of a human body (usually a speaker) and a background. The conveyed visual information of primary interests in videophone communication is usually the human bodies in the scene.

Block-based motion compensation has been widely adopted by the prevalent video coding standards, such as H.261/263 and MPEG 1/2/4. It is based on the translational rigid motion model of blocks, and its popularity is due to the low computational complexity and the low overhead requirements to represent the motion field. However, the translational rigid motion model fails for zooming, rotational motion, and deformations of nonrigid objects (such as face and hand motion [1]). As a nonrigid object, the human body involves complex motion, rotation (of hands and the head) and local deformations. If the translational rigid motion model of blocks is applied to videophone coding, it results in poorer coding quality for the foreground region than the background region, as will be illustrated in Section II. Therefore it is desirable to devise a perceptual model to differentiate the foreground and background regions for rate control so that the available bits can be assigned to maximize the perceptual quality of coded video.

Several schemes [2]–[5] have been proposed recently to improve the performance of coding systems in terms of perceptual quality of foreground objects, by exploiting certain prior knowledge in videophone applications. These schemes firstly segment facial regions from the scene, and then apply finer quantization in facial region and coarser quantization in nonfacial region. However, the existing schemes lack a unified perceptual importance model for rate control. Consequently the bit-rates between facial region and nonfacial region are heuristically controlled, and the quantization schemes cannot be adaptive to local perceptual significance within the facial region and the nonfacial region.

There are two categories of factors affecting the perception of the human visual system (HVS) [6], [7]: stimulus-driven factors and cognition-driven factors. The former category refers to the factors relating to the optical property of eyes and the retina structure, as well as the signal transmission property via the vision path. There are various such factors that could be modeled into video quality assessment, e.g., color, spatial masking, temporal masking, and motion. Considering the low-delay, low-power and low-resolution constraints of the videophone application, we only use two spatial masking related factors in this study: luminance adaptation and texture masking. Luminance adaptation refers to the fact that the HVS is sensitive to luminance contrast rather than absolute luminance value, while texture masking denotes that textured regions can hide more error than smooth areas. In our previous work [8], a nonlinear additivity model for masking (NAMM) has been proposed to combine luminance adaptation and texture masking effects into a just noticeable distortion (JND) profile of image.

Cognition-driven factors reflects the human's cognitive processing, such as object/pattern recognition based on the knowledge and experience. In the videophone application, the presence of the human body no doubt is a cognitive-driven factor attracting visual attention. In general, the face is the most significant part of visual attention in the conversational (usually head-and-shoulder) application; however, in some scenario such as the *Silent* test sequence where hand-language is involved, hands are even more important than the face. Skin color can

Manuscript received August 11, 2003; revised March 2, 2004. This paper was recommended by Associate Editor R. Lancini.

X. Yang was with Infocomm Research, 119613, Singapore. He is now with the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: xkyang@ieee.org).

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

Digital Object Identifier 10.1109/TCSVT.2005.844458

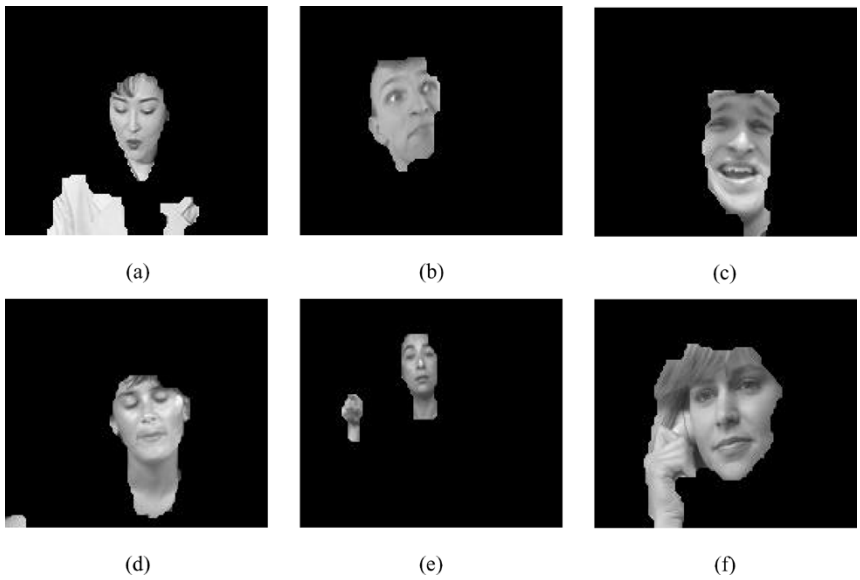


Fig. 1. Skin maps for six standard testing sequences at the 120th frame. (a) Akiyo. (b) Carphone. (c) Foreman. (d) Miss America. (e) Silent. (f) Suzie.

be used as a cognition-driven perceptual cue, since both faces and hands are normally present in the regions with skin color. Face detection is currently still a challenging task because of the variability in scale, location, orientation, and pose [9]. If false face detection occurs from time to time in a video sequence and if such face maps were used for rate control, the perceptual quality would be greatly jeopardized due to the quality variation in human faces. The detection of skin color is much easier and more robust.

In this paper, we present a method for formulating the local perceptual cues and integrating them for both foreground and background objects. The proposed scheme is then applied to rate control in videophone. The remainder of this paper is organized as follows. In Section II, we evaluate the typical block-based coding scheme for the videophony application to demonstrate the necessity of a new rate control scheme that takes the unequal perceptual importance among different visual objects into account. In Section III, we present the perceptual sensitivity by extracting and integrating luminance adaptation, texture masking and skin color clues, as well as its application for the perceptual rate control (PRC) in videophone. In Section IV, both objective and subjective quality evaluations are given to compare the proposed PRC method with the H.263 TMN8 rate control scheme [10]. Finally, conclusions are drawn in Section V.

II. Analysis of Block-Based Coding for Videophony

Due to its simplicity, block-based coding has been widely adopted by the prevalent video compression standards, so is the primary choice for videophony. We now examine the peak signal-to-noise ratio (PSNR) characteristics of block-based coding with videophony-like sequences. Let PSNR_F, PSNR_B, and PSNR_W denote the PSNR of the foreground region, the background region and the whole image, respectively.

We use six typical videophony-like scenes from standard test sequences: *Akiyo*, *Carphone*, *Foreman*, *Miss America*, *Silent*, and *Suzie*. These six original sequences of 30 f/s are compressed into 10 f/s using an H.263 encoder with the TMN8 rate control scheme to achieve constant bit rate (CBR) while maintaining a low buffer delay [10]. In our experiments, the first frame is intracoded (I-frame) with a fixed quantizer. By default, we use $QP = 13$, corresponding to a step size of 26. The following frames are intercoded (P-frame), i.e., they are predicted by the respective previous encoded frames using motion compensation. Encoding only P-frames (after an I-frame) is a common strategy for keeping the end-to-end delay low in video communications. The foreground region is detected using the skin color detection algorithm (to be described in Section III-B), and the detection results are illustrated in Fig. 1.

The average PSNR_F, PSNR_B and PSNR_W of the six coded sequences are presented in Fig. 2, for a wide range of bit rates. It can be seen that PSNR_F is consistently lower than the PSNR_B for all sequences. This implies that foreground objects need more bits with block-based coding due to nonrigid deformations and the underlying motion of more complex nature (i.e., rotation). The conventional video coding scheme without considering unequal perceptual importance in rate control results in lower PSNR for the foreground. Such results are undesirable because of the inconsistence with the HVS perception. Therefore, it is necessary to develop a new rate control scheme that takes the unequal perceptual importance among different objects into account.

III. Local Perceptual Cues and the Use in Videophone Rate Control

Fig. 3 illustrates a flow chart of the proposed scheme to extract and integrate local perceptual cues for macroblock-level rate control of videophone. There are four major modules in Fig. 3: stimulus-driven sensitivity detection, skin color detection, perceptual integration, and postprocessing. In current implementation, luminance adaption and texture masking are extracted as stimulus-driven factors, while skin color is used as the cognition-driven factor. Then, the results of stimulus-driven sensitivity and skin color detection are integrated into a cognition-integrated perceptual sensitivity map. After the

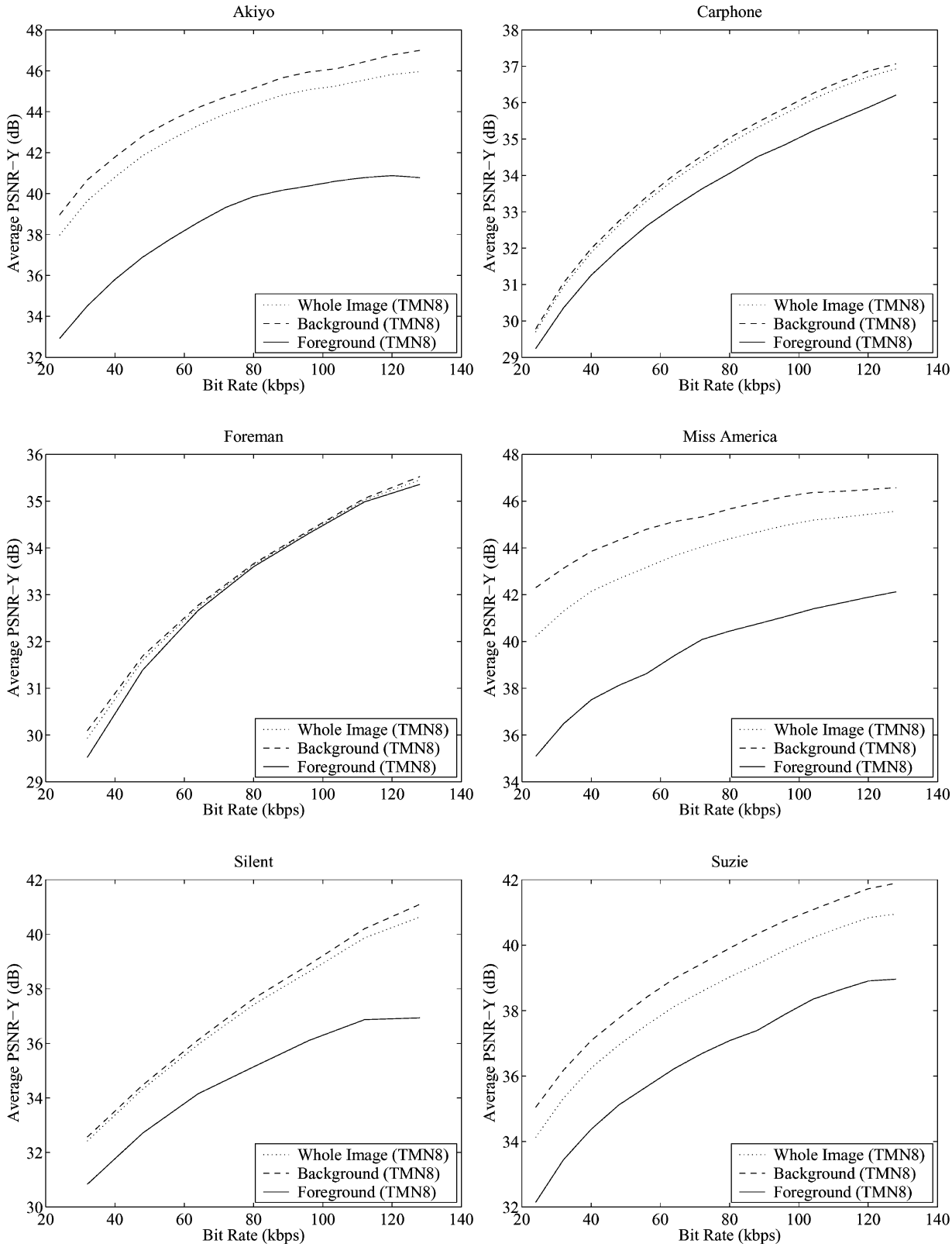


Fig. 2. PSNR of foreground, background, and whole image for six sequences compressed by H.263 TMN8 coder.

postprocessing, the resultant perceptual weight for each macroblock is formed to guide the determination of quantization step with the rate-distortion model, and allows fully adaptive bit allocation for both foreground and background objects. The outputs of the four modules for the 120th frame of *Carphone* sequence are also shown in Fig. 3.

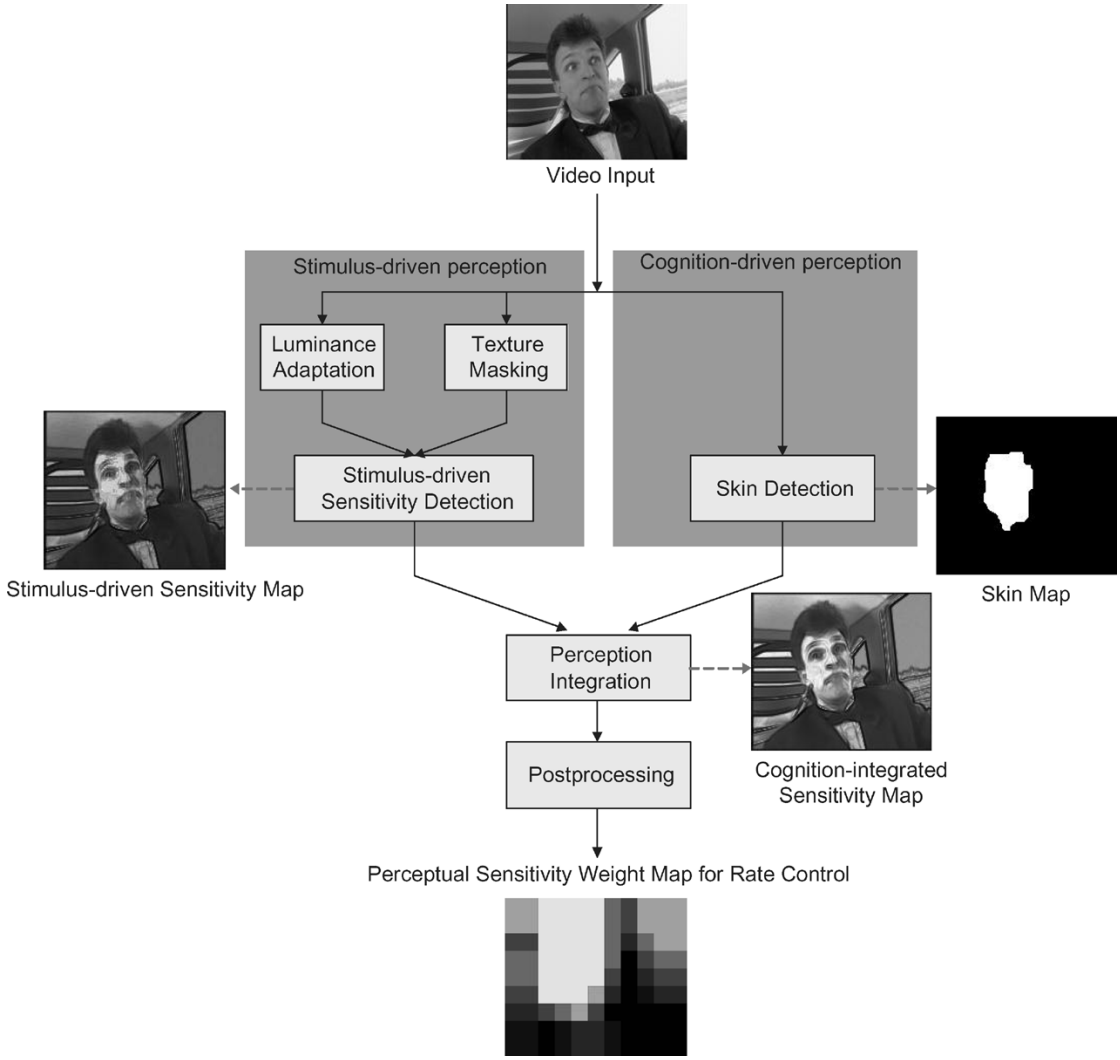


Fig. 3. Flow chart for extracting and integrating perceptual cues.

A. Stimulus-Driven Sensitivity

In this paper, the stimulus-driven sensitivity for a pixel in an image is defined by combining the luminance adaptation and texture masking effects (as the reciprocal of JND [8]). Let $I(x,y)$ denote the luminance intensity for the pixel located at (x,y). The associated stimulus-driven sensitivity can be given by

$$S(x,y) = (T_l(x,y) + T_t(x,y) - C_{l,t} \cdot \min\{T_l(x,y), T_t(x,y)\})^{-1} \quad (1)$$

where $T_l(x,y)$ and $T_t(x,y)$ are the visibility thresholds for luminance adaptation and texture masking; $C_{l,t}(\in (0,1))$ is the gain reduction factor due to overlapping between two masking stimuli.

$T_l(x,y)$ can be determined as in [11] for 8-bit image representation

$$T_l(x,y) = \begin{cases} 17\left(1 - \sqrt{\frac{\overline{I}(x,y)}{127}}\right) + 3, & \text{if } \overline{I}(x,y) \le 127 \\ \frac{3}{128}\left(\overline{I}(x,y) - 127\right) + 3, & \text{otherwise} \end{cases} \quad (2)$$

1	1	1	1	1
1	2	2	2	1
1	2	0	2	1
1	2	2	2	1
1	1	1	1	1

Fig. 4. Weighted low-pass filter for calculating average background luminance.

where $\overline{I}(x,y)$ is the average of $I(x,y)$, calculated by a 5×5 weighted low-pass filter as shown in Fig. 4.

$T_t(x,y)$ can be computed as follows:

$$T_t(x,y) = \beta \cdot G(x,y) \quad (3)$$

where $\beta(= 0.117)$ is an empirical parameter and is determined according to our extensive subjective tests with the viewing distance of approximately six times of the image height, using a 21″ EIZO T965 professional color monitor with 1600×1200 resolution; $G(x,y)$ denotes the maximum of the four weighted local gradients computed by four 5×5 directional high-pass filters as shown in Fig. 5.

0	0	0	0	0
1	3	8	3	1
0	0	0	0	0
-1	-3	-8	-3	-1
0	0	0	0	0

0	0	1	0	0
0	8	3	0	0
1	3	0	-3	-1
0	0	-3	-8	0
0	0	-1	0	0

0	0	1	0	0
0	0	3	8	0
-1	-3	0	3	1
0	-8	-3	0	0
0	0	-1	0	0

0	1	0	-1	0
0	3	0	-3	0
0	8	0	-8	0
0	3	0	-3	0
0	1	0	-1	0

Fig. 5. Directional high-pass filters for texture detection.

(a) Original

(b) PSNR=35.58 dB

(c) PSNR=35.53 dB

Fig. 6. The 120th frame of *Carphone*. (a) Original image. (b) Corrupted by random noise injection. (c) Corrupted by noise injection with the cognition-integrated sensitivity map.

The stimulus-driven sensitivity map of *Carphone* is given in Fig. 3, where the brightness represents the sensitivity strength (i.e., the brighter, the larger stimulus-driven sensitivity).

B. Skin Color Detection

Human skin color has been proven to be an effective feature in many applications from face detection to hand tracking [9]. Several color spaces have been utilized to label pixels as skin [9]. YC_bC_r color space is adopted due to: 1) YC_bC_r color space has been used in the prevalent image/video compression standards and 2) YC_bC_r color space leads to good performance of skin color detection in terms of the separation between luminance and chrominance [13], and the compactness of the skin clusters [4], [14].

Although in many studies it is assumed that skin-tone color is independent of luminance component [4], [14], in practice, the skin-tone color is nonlinearly dependent on luminance, especially on the extreme luminance [15], [16]. The difficulty of detecting the low-luma and high-luma skin tones can be efficiently overcome by applying nonlinear transform in YC_bC_r color space. Hsu's nonlinear transform [15] of chroma in YC_bC_r color space is used in this paper.

With nonlinear transform, the chroma (C_b and C_r) are transformed as the functions of Y: $C'_b(Y)$ and $C'_r(Y)$ for every pixel. The elliptical skin model for the skin tones in the transformed $C'_bC'_r$ space is described by

$$\frac{(x-e_{c_x})^2}{a^2}+\frac{(y-e_{c_y})^2}{b^2}=1 \tag{4}$$

$$\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix}\begin{bmatrix} C'_b-c_x \\ C'_r-c_y \end{bmatrix} \tag{5}$$

where $c_x=109.38$, $c_y=152.02$, $\theta=2.53$ (in radian), $e_{c_x}=1.60$, $e_{c_y}=2.41$, $a=25.39$, and $b=14.03$.

Fig. 1 shows the detected skin maps for six standard testing sequences at the 120th frame. The rough segmentation for skin regions provides effective indication on the whereabouts of foreground objects. As mentioned in the Introduction, skin region

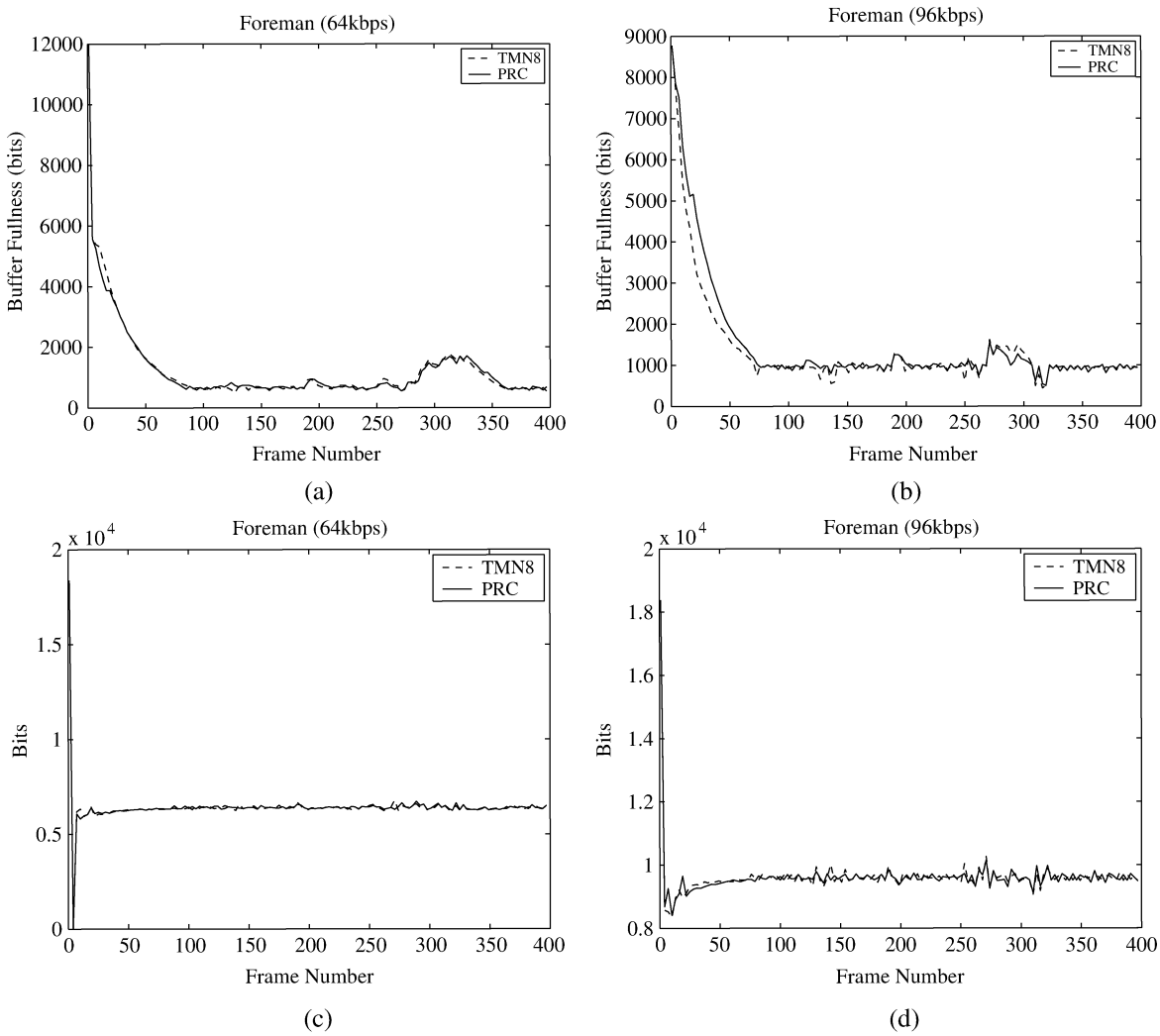


Fig. 7. Comparison of the buffer fullness and the bits per frame when the proposed PRC (solid line) and TMN8 rate control (dashed line) are used in the H.263 codec. (a) Buffer fullness for *Foreman* at 64 kb/s. (b) Buffer fullness for *Foreman* at 96 kb/s. (c) Bits per frame for *Foreman* at 64 kb/s. (d) Bits per frame *Foreman* at 96 kb/s.

detection is a more meaningful and reliable approach for foreground detection in the videophone application than face detection.

C. Perceptual Integration

Since the distortion in the skin region of videophone is the most intolerable from the viewpoint of cognition-driven perception, we compute the cognition-integrated perceptual sensitivity by simply scaling $S(x,y)$ values for the pixels within the skin map $\mathcal{F}$ as follows:

$$S_c(x,y) = \begin{cases} \dfrac{\max\limits_{\substack{1\le x\le X\\ 1\le y\le Y}}(S(x,y))}{\max\limits_{(x,y)\in\mathcal{F}}(S(x,y))} \cdot S(x,y), & \text{if } (x,y)\in\mathcal{F} \\ S(x,y), & \text{otherwise} \end{cases} \quad (6)$$

where X and Y denote the image dimensions.

Fig. 6 illustrates how the $S_c(x,y)$ profile reflects the perceptual sensitivity in an image. In Fig. 6, the original image of the 120th frame of *Carphone* sequence [Fig. 6(a)] is corrupted by random noise injection as in Fig. 6(b). As can be seen, the perceptual quality of the randomly noised image is very poor. If we shape the same amount of noise (therefore, with similar PSNR) according to the $S_c(x,y)$ profile (i.e., more noise in less sensitive areas), the noise is almost invisible [as in Fig. 6(c)]. $S_c(x,y)$ reflects the local perceptual sensitivity for coding error and, therefore, is the basis of sensible bit allocation for both foreground and background objects.

D. Postprocessing

In typical block-based video coders, the number of the bits used and the resultant distortion for a given macroblock depend on the macroblock's quantization parameter (QP), which determines the step size Q for quantizing the transformed coefficients. We can determine a perceptual sensitivity weight for the nth macroblock ($1 \le n \le N$, and $N = XY/256$), denoted as $w_o(n)$, as follows:

$$w_o(n) = \frac{\sum\limits_{i=1}^{16}\sum\limits_{j=1}^{16} S_c(n,i,j)}{A} \quad (7)$$

where $S_c(n,i,j)$ represents the cognition-integrated sensitivity at the (i,j)th pixel of the nth macroblock, and A is the number

TABLE I
COMPARISON OF TMN8 AND PRC

Sequences	Total frames in the sequence	Bitrate (kbps)	no. of skipped frames		DMOS	
			TMN8	PRC	TMN8	Difference with PRC
Akiyo	300	24	6	6	45.2	-9.4
		48	2	2	27	-12
Carphone	350	64	1	1	64	-17
		96	0	0	48	-15
Foreman	400	64	1	1	70	-15
		96	0	0	51	-18.4
Miss America	150	24	2	2	46	-17
		48	0	0	30	-13
Silent	400	48	2	2	52	-6
		64	1	1	40	-9
Suzie	150	48	1	1	68	-7
		64	0	0	52	-9
Mean DMOS difference between PRC and TMN8 for all sequence						-12.3

of pixels in a macroblock (i.e., $A = 16 \times 16$). The initial map of the sensitivity weights is: $\mathbf{w}_o = \{w_o(n)\}$.

In H.263 and MPEG-4 coding, QPs are differentially encoded in a raster-scan order, and there is, on average, a 5-bit penalty for changing the quantizer value. Specifically, 2 bits are used in the syntax element DQUAT to indicate the change of value for QP (for H.263, DQUANT, and QP are restricted to values in $\{-2, -1, 1, 2\}$ and $\{1, 2, \ldots, 31\}$, respectively), and three additional bits are needed on average in the syntax element MCBPC to indicate the change in the current block against the previous macroblock. At a high bit rate, these overheads resulting from changing the quantizer are negligible, but with the decrease of the bit rate, the ratio of these overheads to other bits becomes increasingly significant. To avoid too frequent changes in QP, $\mathbf{w}_o$ is filtered with the following 3×3 gray-scale morphological closing operator to yield a modified weight map which is not changed too frequently in a neighborhood

$$\mathbf{w} = \mathbf{w}_o \bullet b = (\mathbf{w}_o \oplus b) \ominus b \tag{8}$$

where $\mathbf{w}_o \bullet b$ denotes the closing of $\mathbf{w}_o$ by 3×3 structuring element b, $\oplus$ and $\ominus$ are dilation and erosion operators, respectively [17], and the weight map after post-processing is $\mathbf{w} = \{w(n)\}$.

The postprocessing presented above also reduces the influence of the inaccuracy of skin region detection toward the final sensitivity weights. The resultant $\mathbf{w}$ for *Carphone* is given as the last part of Fig. 3, where the brightness represents the strength of perceptual sensitivity. It can be seen that the formation of $\mathbf{w}$ is in line with the HVS perception.

E. PRC

In video phone situations, the distortion for the nth macroblock is mainly introduced by quantizing its DCT coefficients. With the rate-distortion model [18], the number of bits invested in the nth macroblock $B(n)$ and the weighted distortion $D(n)$, can be, respectively, determined by

$$B(n) = A\left(K\frac{\sigma^2(n)}{Q^2(n)} + C\right) \tag{9}$$

$$D(n) = w^2(n) \cdot \frac{Q^2(n)}{12} \tag{10}$$

where $Q(n)$ is the quantization step size, K and C are constants, and $\sigma(n)$ is the standard deviation of the luminance and chrominance values in the macroblock. If $w(1) = w(2) =, \ldots, w(N)$ (i.e., the HVS characteristics are not considered), $D(n)$ is approximately the mean squared error (MSE) between the original and the encoded macroblocks.

The optimized quantization step size $Q^*(n)$ can be determined by minimizing the overall distortion $D = \sum_{n=1}^{N} D(n)/N$ subject to a given bit allocation $B = \sum_{n=1}^{N} B(n)$. This constrained optimization problem can be solved by Lagrangian optimization [18]

$$Q^*(n) = \sqrt{\frac{A \cdot K \cdot \sigma(n)}{(B - A \cdot N \cdot C) \cdot w(n)} \sum_{n=1}^{N} w(n)\sigma(n)}. \tag{11}$$

Basically, the proposed PRC introduces the perceptual sensitivity weights $w(n)$ into the CBR rate controller for bit assignment at macroblock-level. It adopts the frame-level control method in TMN8 and therefore inherits the accuracy for bit assignment at frame-level (as will be shown in Fig. 7). The impact of the perceptual bit allocation is at macroblock-level: the proposed PRC scheme optimizes the use of the allocated bits/frame for maximum perceptual benefit (guided by $w(n)$). However, the proposed method at macroblock-level control does not separate the allocation between foreground and background; actually, the unequal bit allocation between foreground and background is achieved by a single process guided with $w(n)$, as aforementioned.

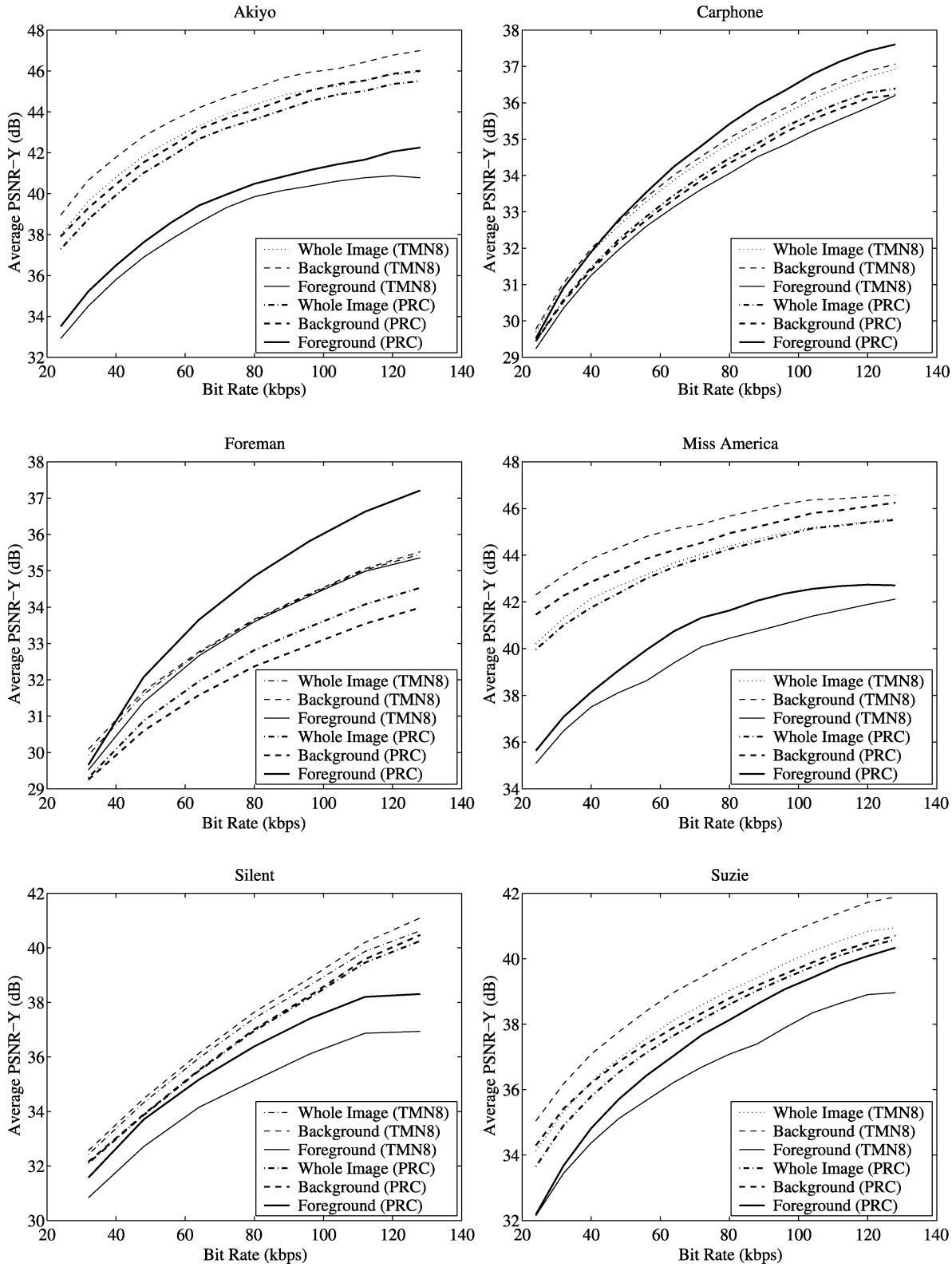


Fig. 8. PSNR comparison between TMN8 rate control scheme and the proposed PRC scheme for foreground, background, and whole image in six sequences.

In the proposed solution, additional computational complexity is involved in the calculation of perceptual sensitivity weights $w(n)$, and is about 8% of the computational complexity of full-search motion estimation with a typical searching window $(-16, 16)$.

IV. Performance Evaluation

The proposed PRC scheme is implemented into the H.263 TMN8 [10] coder. The aforementioned six test sequences, whose skin maps are exemplified in Fig. 1, are coded at 10 f/s.

(a) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{32.18}, 31.73, 31.79)dB$

(b) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{33.25}, 30.69, 30.97)dB$

(c) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{33.30}, 33.42, 33.41)dB$

(d) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{35.25}, 32.21, 32.52)dB$

(e) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{34.08}, 34.55, 34.48)dB$

(f) $(PSNR_F, PSNR_B, PSNR_W) = (\mathbf{36.75}, 33.35, 33.69)dB$

Fig. 9. The 120th decoded frame for *Foreman* compressed by: (a) TMN8 at 64 kb/s, (b) PRC at 64 kb/s, (c) TMN8 at 96 kb/s, (d) PRC at 96 kb/s, (e) TMN8 at 128 kb/s, and (f) PRC at 128 kb/s.

A. Objective Quality Evaluation

Fig. 7(a)–(b) plots the buffer fullness while (c)–(d) shows the used bits per frame throughout the encoding process, with the proposed PRC (solid line) and the TMN8 rate control (dashed line) for *Foreman* sequence (the results are almost the same for the other sequences). From the figure, we can see that the traces of both the buffer fullness and the used bits per frame are very similar for the two schemes. The first part of Table I further compares the number of skipped frames for different sequences. It can be concluded from Fig. 7 and Table I that the proposed PRC and the TMN8 rate control have very similar behavior in

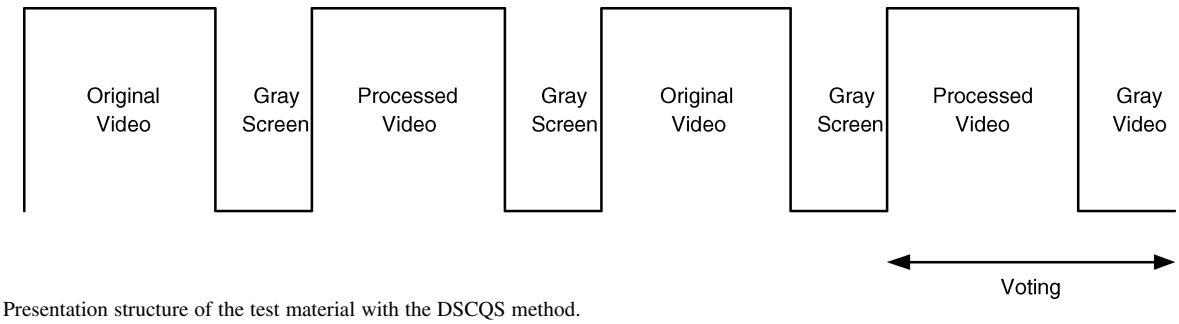


Fig. 10. Presentation structure of the test material with the DSCQS method.

terms of the buffer fullness, the used bits per frame, and the number of skipped frames. Therefore the proposed PRC scheme demonstrates how to assign the allocated bits within each frame (approximately the same number of bits as the TMN8 scheme) according to the perceptual cues toward enhanced visual quality.

Fig. 8 compares the average PSNRs between the TMN8 rate control and the proposed PRC schemes for foreground, background and whole image in the six sequences with a range of bit rates. Similar to the experiments in Section II, the foreground is indicated by the skin region. It can be seen that PSNR_F is significantly improved by the PRC scheme in comparison with the TMN8 rate control scheme. For *Carphone* and *Foreman* sequences, PSNR_F becomes higher than PSNR_B after the proposed PRC scheme is incorporated. For the other four sequences, PSNR_F is still lower than PSNR_B after the proposed PRC scheme is incorporated, but the gap between PSNR_F and PSNR_B is greatly reduced. We can see that the increase of PSNR_F is at the cost of the slight decrease of PSNR_B and PSNR_W; however, as will be confirmed by the subjective viewing tests in the next subsection, the quality increase of foreground objects brings about significant overall improvement in perceptual quality at a given bit rate.

Fig. 9 presents the snapshots of the 120th decoded frame for *Foreman* compressed by TMN8 and the proposed PRC scheme, respectively, at various bit rates. Both blurring and blocking effects in the region with skin color are greatly alleviated by the proposed PRC scheme, while the quality difference in the background are almost unnoticeable.

B. Subjective Quality Evaluation

Double stimulus continuous quality scale (DSCQS) method, proposed by Rec. ITU-R BT.500 [19], was used to evaluate the subjective quality of a decoded sequence (obtained with the proposed PRC or the H.263 TMN8 rate control scheme) relative to its original sequence. Each display and voting session for an original sequence and an associated decoded sequence is illustrated in Fig. 10. Both the display order of the sequences in a session and the order of the six test sequences were randomized for viewers. The mean opinion score (MOS) scales for viewers to vote for the quality after viewing are: excellent (100-80), good (80-60), fair (60 -40), poor (40-20) and bad (20-0). Ten observers were involved in the experiments. The subjective visual quality assessment was performed in a typical laboratory environment with normal lighting, using a $21''$ EIZO T965 professional color monitor with resolution of 1600×1200. The viewing distance is approximately six times of the image height.

Difference mean opinion scores (DMOS) are calculated as the difference of MOSs between the original video and the decoded video. The smaller the DMOS is, the higher perceptual quality of the decoded video has when compared with the original video. Table I compares the averaged DMOSs over the all ten observers for the six sequences. It shows that the DMOS is consistently reduced by the proposed PRC for each case (i.e., the subjective rating is consistently better for the decoded sequences with the PRC), and an average subjective quality gain of 12.3 measured in DMOS is achieved by the proposed scheme.

V. Conclusion

In videophone/videoconferencing, the block-based video coding schemes tend to result in lower coding quality with foreground objects (e.g., the head and hands of the talking person(s)) than background objects. This is because of the poorer motion estimation as the result of nontranslational motion and the nonrigid deformations of human bodies in the scene. Hence, it is desirable to allocate bits effectively throughout the frame according to the HVS sensitivity for quality perception, since the available bandwidth is very limited in such an application.

In this paper, local perceptual cues have been extracted from the visual signal. For efficiency, luminance adaption and texture masking are extracted as stimulus-driven factors, while skin color is used as the cognition-driven factor. The resultant perceptual sensitivity weight for each macroblock is formed to guide the determination of quantization step with the rate-distortion model, and allows adaptive bit-allocation for both foreground and background objects. It has been demonstrated that in comparison with the existing H.263 TMN8 rate control scheme the proposed PRC scheme achieves higher PSNR for foreground objects at a fixed bit rate. The subjective viewing tests further confirm the overall perceptual quality improvement by the PRC scheme.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable advice toward the improvement of this paper.

References

[1] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ: Prentice-Hall, 1995.

[2] J. Hartung, A. Jacquin, J. Pawlyk, J. Rosenberg, H. Okada, and P. E. Crouch, "Object-oriented h.263 compatible video coding platform for conferencing applications," *IEEE J. Sel. Areas Comm.*, vol. 1, no. 3, pp. 264–277, Mar. 1999.

[3] S. Daly, K. Matthews, and J. Ribas-Carbera, "Face-based visually-optimized image sequence coding," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, Chicago, IL, Oct. 1998, pp. 443–447.
[4] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone application," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 6, pp. 551–564, Jun. 1999.
[5] C.-W. Lin, T.-J. Liao, and Y.-C. Chen, "Low-complexity face-assisted video coding," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Sep. 2000, pp. 207–210.
[6] 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.
[7] Z. K. Lu, W. S. Lin, E. P. Ong, S. S. Yao, and X. K. Yang, "Perceputal-quality significance map (PQSM) and its application on video quality distortion metrics," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, Hong Kong, Apr. 2003, pp. 617–620.
[8] 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 *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, Hong Kong, Apr. 2003, pp. 609–612.
[9] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.
[10] *Video Codec Test Model, Near-Term, Version 8 (TMN8)*, ITU-T/SG15, Jun. 1997.
[11] 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. 6, pp. 467–476, Jun. 1995.
[12] 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.
[13] M. J. Nadenau, "Integration of Human Color Vision Models Into High Quality Image Compression," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2000.
[14] H. Wang and S. F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 615–628, Apr. 1997.
[15] R.-L. Hsu, A.-M. Moharmed, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 5, pp. 696–706, May 2002.
[16] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 16, pp. 42–54, Jan. 1998.
[17] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Reading, MA: Addison-Wesley, 1992.
[18] J. Ribas-Corbera and S. Lei, "Rate control in dct video coding for low-delay communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 172–185, Feb. 1999.
[19] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R REC. BT. 500-9, 1999.

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

He is currently an Associate Professor in the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University, China. From April 2002 to October 2004, he was a Research Scientist in the Institute for Infocomm Research, Singapore. From September 2000 to March 2002, he worked as a Research Fellow in Centre for Signal Processing, Nanyang Technological University, Singapore. He has published over 50 refereed papers. His current research interests include scalable video coding, video transmission over networks, video quality assessment, digital television, and pattern recognition.

Dr. Yang received the Best Young Investigator Paper Award at SPIE International Conference on Video Communication and Image Processing (VCIP2003). He is currently a member of Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society.

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

He had taught and/or researched in Zhongshan University, Guangzhou, China, Shantou University, Shantou, China, Bath University, Bath, U.K., and the National University of Singapore and Institute of Microelectronics, Singapore. He has been the project leader of a number of successfully delivered projects in development of digital multimedia related technologies since 1997. He is currently an Associate Lead Scientist in Institute for Infocomm Research, Singapore. His current research interests include perceptual visual distortion metrics, perceptual video coding, and multimedia signal processing.

Zhongkang Lu (S'95–M'99) 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 in the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University, Hong Kong. From 1999 to 2001, he was a Research Fellow in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Currently, he is a Research Scientist in the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include perceptual visual signal processing, pattern recognition and computer vision.

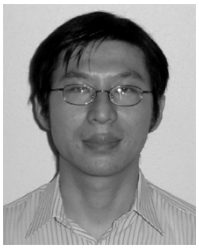
Xiao Lin (SM'02) received the Ph.D degree from the Electronics and Computer Science Department, University of Southampton, Southampton, U.K. in 1993.

He worked with the Centre for Signal Processing (CSP) for about five years as a researcher and manager on Multimedia Program. He worked for DeSOC technology as a Technical Director, where he contributed on the VoIP solution, speech packet lost concealment for Bluetooth, WCDMA baseband SOC development. He joined Institute for Infocomm Research, Singapore in 2002, where he is now Research MANAGER in charge of multimedia signal processing areas.

Susanto Rahardja (SM'03) received B.Eng. degree in electrical engineering in 1991 from the National University of Singapore (NUS), Singapore, the M.Eng. degree in digital communication and microwave circuits, and the Ph.D. degree in the area of logic synthesis and signal processing from the Nanyang Technological University (NTU), Singapore, in 1993 and 1997, respectively.

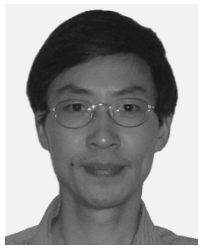
He joined the Centre for Signal Processing, NTU, as a Research Engineer in 1996, a Research Fellow in 1997, and served as a business development manager in 1998. In 2001, he joined NTU as an Academic Professor and was appointed the Assistant Director of the Centre for Signal Processing. In 2002, he joined the Agency for Science, Technology and Research and was appointed as the Program Director to lead the Signal Processing program. He is currently the director of Media Division in the Institute for Infocomm Research. He has more than 100 articles in international journals and conferences. He is currently an Associate Professor at the School of Electrical and Electronic Engineering in the Nanyang Technological University. His research interests include binary and multiple-valued logic synthesis, digital communication systems, and digital signal processing.

Dr. Rahardja was the recipient of the prestigious IEE Hartree Premium Award in 2002 and the Tan Kah Kee Young Inventors' GOLD Award (Open Category) in 2003. He is the cofounder of AMIK Raharja Informatika and STMIK Raharja, an institute of higher learning in Tangerang, Indonesia.



EePing Ong (M'02) 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. 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 Research Scientist. His research interests include optical flow, motion estimation, video object segmentation and tracking, perceptual image/video quality metrics, perceptual image/video coding, and multimedia signal processing.



Susu Yao (M'97) received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in July 1993.

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