JOINT ADAPTIVE BACKGROUND SKIPPING AND WEIGHTED BIT ALLOCATION FOR WIRELESS VIDEO TELEPHONY

Haohong Wang and Khaled El-Maleh

Qualcomm Incorporated
San Diego, CA 92121, USA
Email: {haohongw, kelmaleh}@qualcomm.com

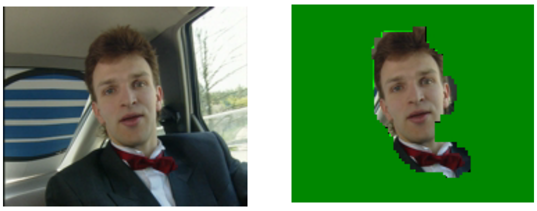
ABSTRACT

In this paper, we propose a novel region-of-interest (ROI) video coding algorithm for wireless video telephony applications. In order to improve the visual quality of the ROI, the proposed approach reallocates bits from Non-ROI macroblocks to ROI by adaptively skipping Non-ROI and using an optimized weighted bit allocation scheme to bias the bit allocation. To the best of our knowledge, this work is the first effort to develop an optimized ρ-domain bit allocation scheme for ROI video coding. Experimental results indicate that the proposed approach significantly outperforms other methods by up to 2dB.

1. INTRODUCTION

Region-of-Interest (ROI) video coding [1-11] has recently become a very attractive field with the increasing popularity of wireless video telephony applications. For these applications with the constraint of very low-bitrate (less than 64kbps), ROI coding can effectively improve the subjective quality of the encoded video sequence by coding certain regions, such as facial image region, at higher quality. As shown in Fig. 1, the facial region in Fig. 1(b) is normally considered as an ROI for the original image shown in Fig. 1(a).

In video telephony applications, the mechanism to accurately specify the ROI is not a trivial task, which has resulted in advanced research on face detection, face segmentation, and object tracking. The current study on ROI video coding generally considers head-and-shoulder video sequences and assumes the ROI (human faces) can be detected in acceptable accuracy. Therefore, bit allocation has always been one of the major focuses of ROI video coding. In [1-2], a face model is used to assist encoding eyes, mouth and other areas with different quantization parameters and temporal resolution. In [3-4], human visual sensitivity variations with eccentricity has been considered in macroblock-level quantizer assignment, where the viewers are assumed most likely gaze at faces, and therefore the visual sensitivities of macroblocks will decrease with distance from the center of gaze. In [5], a simple two-level quantization scheme is used which assigns a finer quantizer to the foreground and a coarser quantizer to the background. In [6-7], spatial or temporal filters are used on background to decrease the bitrate cost and save bits for the foreground. In [8], the macroblocks are classified into three regions: face region, active non-face region and static non-face region. The encoding for the static non-face macroblocks are skipped to save bits for compensating the quality of the face and active non-face macroblocks.



(a) Original Image (b) ROI of the image

Figure 1 An example of ROI

So far, most of the current ROI bit allocation algorithms [7-11] are based on a weighted version of the H.263+ TMN8 model [12], where a cost function is created and the distortion components on various regions in the function are treated differently by using a set of preset weights. Like most of the other video standards, TMN8 uses a Q-domain rate control scheme, which models the rate and distortion with functions of quantization step size (QP). However, recent advances in rate control research and development have demonstrated that the ρ-domain rate control model [13] (ρ represents the number (or percentage) of non-zero quantized AC coefficients in a macroblock in video coding) is more accurate and thus effectively reduces the rate fluctuations. Our simulations also showed that the ρ-domain rate control significantly outperformed the Q-domain approach. To the best of our knowledge, so far there is not a general optimized ρ-domain bit allocation model for ROI video coding, although an ad-hoc bit allocation solution in the ρ-domain was used in [6].

In this paper, we propose a ρ-domain optimized weighted bit allocation scheme for ROI video coding. At the same time, we develop an adaptive background skipping algorithm, which can be used jointly with the weighted bit allocation scheme. The major difference between the proposed skipping approach and the macroblock skipping feature of [8] is that the proposed method can dynamically control background skipping based on the statistics of the previous coded frame. In other words, our proposed method will turn off background skipping if it severely hurts the video fidelity, for example when the background content contains important information.

The rest of the paper is organized as follows. Section 2 presents the optimized weighted bit allocation scheme, Section 3 proposes the adaptive background skipping approach, and Section 4 demonstrates the experimental results. We draw conclusions in the last section.

2. OPTIMIZED WEIGHTED BIT ALLOCATION

Although ROI video coding has become a very popular research topic recently, quality measurement for ROI video is still an open issue. Most papers use PSNR as a distortion metric to evaluate the quality of ROI and Non-ROI, however, the overall video quality cannot be accurately evaluated. In the following text, we propose a perceptual quality measurement for ROI video. It is important to point out that the main purpose of this paper is not to develop a general quality metric for ROI video, instead we introduce a quality cost function to bias the bit allocation scheme into a subjective visual favorable solution. Three major factors should be considered in the metric: users' interest, video fidelity and perceptual quality of the reconstructed video data. The users' interest directly determines the classification of a video frame into ROI and Non-ROI parts and their associated perceptual importance factors. In video telephony applications, speaker's face region is a typical ROI because the human being's facial expression is very complicated and small variation can convey large quantity of information. For video fidelity factor, PSNR is a good measurement, which indicates the total amount of distortion of the reconstructed video frame compared to the original frame. In general, fidelity is the most important consideration for video coding, where any improvement might cause better subjective visual quality. However, it is not always the case, and that is why perceptual quality factors should also be taken into account. The perceptual quality considers both spatial errors, for example blocking and ringing artifacts, and temporal errors such as temporal flicker where the frame visual qualities change non-uniformly along the temporal axis. In this work, we only consider the first two factors in the metric in our design of real-time video communication systems.

Let us denote by D_R and D_{NR} the normalized per pixel distortion of the ROI and Non-ROI, and α the ROI perceptual importance factor. If we assume the relationship among the aspects mentioned above can be simplified into a linear function in video quality evaluation, then we can represent the overall distortion of a video frame as

$$D_{Frame} = \alpha D_R(f,\tilde{f}) + (1-\alpha) D_{NR}(f,\tilde{f}), \qquad (1)$$

where f and $\tilde{f}$ are the original and reconstructed frames, D_R and D_{NR} are the normalized errors of ROI and Non-ROI in fidelity. It is clear that α should be assigned real values between 0 and 1, and the selection of α is up to end-users based on their requirements and expectations. Again, this measurement is not a perfect metric, but it will be shown in the subsequent tests in section 4 to help the bit allocation process to favor subjective perception.

Let us denote by R_{budget} the total bit budget for a given frame f and R the bit rate for coding the frame, then the problem can be represented by

$$\text{Minimize } D_{Frame}\text{, such that } R \le R_{budget}. \qquad (2)$$

Clearly, this optimization problem can be solved by Lagrangian relaxation and dynamic programming in the same fashion as in [14]. However, the computational complexity is tremendously higher than a real-time system can bear. Therefore, we propose a low-complexity near-optimal solution as follows.

In ROI video coding, let us denote by N the number of macroblocks in the frame, $\{\rho_i\}$, $\{\sigma_i\}$, $\{R_i\}$ and $\{D_i\}$ the set of ρ's, standard deviation, bit rate and distortion (sum of squared error) for the ith macroblock. Thus, $R=\sum_{i=1}^{N} R_i$.

We define a set of weights $\{w_i\}$ for each frame as

$$w_i = \begin{cases} \frac{\alpha}{K} & \text{if it belongs to ROI} \\ \frac{1-\alpha}{(N-K)} & \text{if it belongs to Non-ROI} \end{cases}, \qquad (3)$$

where K is the number of macroblocks within the ROI. Therefore, the weighted distortion of the frame is

$$D = \sum_{i=1}^{N} w_i D_i$$
$$= [\alpha D_R(f,\tilde{f}) + (1-\alpha) D_{NR}(f,\tilde{f})] * 255^2 * 384, \qquad (4)$$

Hence, problem (2) can be rewritten as

$$\text{Minimize } \sum_{i=1}^{N} w_i D_i \text{, such that } \sum_{i=1}^{N} R_i \le R_{budget}. \qquad (5)$$

We propose to solve (5) by using a model-based bit allocation approach. As shown in [15], the distribution of the AC coefficients of a natural image can be best approximated by a Laplacian distribution $p(x)=\frac{\eta}{2}e^{-\eta|x|}$. Therefore in [12], the rate and distortion of the ith macroblock can be modeled in (6) and (7) as functions of ρ,

$$R_i = A\rho_i + B, \tag{6}$$

where A and B are constant modeling parameters, and A can be thought as the average number of bits needed to encode non-zero coefficients and B can be thought as the bits due to non-texture information.

$$D_i = 384\sigma_i^2 e^{-\theta\rho_i/384}, \tag{7}$$

where θ is an unknown constant.

Here, we optimize with respect to ρ_i instead of quantization parameter because we assume that there is an accurate enough ρ -QP mapping table available to generate a decent quantizer from any selected ρ_i . In general, (5) can be solved by using Lagrangian relaxation in which the constrained problem is converted into an unconstrained problem that

$$\underset{\rho_i}{\text{Minimize}}\ J_\lambda = \lambda R + D = \sum_{i=1}^{N}(\lambda R_i + w_i D_i), \tag{8}$$

$$= \sum_{i=1}^{N}[\lambda(A\rho_i + B) + 384 w_i \sigma_i^2 e^{-\theta\rho_i/384}]$$

where λ^* is the solution that enables $\sum_{i=1}^{N} R_i = R_{budget}$. By setting partial derivatives to zero in (8), we obtain the following expression for the optimized ρ_i, that is

$$\text{let } \frac{\partial J_\lambda}{\partial\rho_i} = \frac{\partial\sum_{i=1}^{N}[\lambda(A\rho_i + B) + 384 w_i\sigma_i^2 e^{-\theta\rho_i/384}]}{\partial\rho_i} = 0, \tag{9}$$

which is

$$\lambda A - \theta w_i \sigma_i^2 e^{-\theta\rho_i/384} = 0, \tag{10}$$

so

$$e^{-\theta\rho_i/384} = \frac{\lambda A}{\theta w_i \sigma_i^2}. \tag{11}$$

and

$$\rho_i = \frac{384}{\theta}[\ln(\theta w_i \sigma_i^2) - \ln(\lambda A)]. \tag{12}$$

On the other hand, since

$$R_{budget} = \sum_{i=1}^{N} R_i = \frac{384A}{\theta}\sum_{i=1}^{N}[\ln(\theta w_i\sigma_i^2) - \ln(\lambda A)] + NB, \tag{13}$$

so,

$$\ln(\lambda A) = \frac{1}{N}\sum_{i=1}^{N}\ln(\theta w_i \sigma_i^2) - \frac{\theta}{384NA}(R_{budget} - NB). \tag{14}$$

From (12) and (14), we obtain bit allocation model I:

$$\rho_i = \frac{384}{\theta}[\ln(\theta w_i\sigma_i^2) - \frac{1}{N}\sum_{i=1}^{N}\ln(\theta w_i\sigma_i^2) + \frac{\theta}{384NA}(R_{budget} - NB)]$$

$$= \frac{R_{budget} - NB}{NA} + \frac{384}{\theta}[\ln(\theta w_i\sigma_i^2) - \frac{\sum_{i=1}^{N}\ln(\theta w_i\sigma_i^2)}{N}]$$

$$= \frac{\rho_{budget}}{N} + \frac{384}{\theta}[\ln(\theta w_i\sigma_i^2) - \frac{\sum_{i=1}^{N}\ln(\theta w_i\sigma_i^2)}{N}], \tag{15}$$

where ρ_{budget} is the total ρ budget for the frame.

In the following text, we will derive another ρ-domain bit allocation model. If we assume that we have a uniform quantizer with step size q, then the distortion caused by quantization is given by

$$D(q) = 2\int_0^{0.5q} p(x)x dx + 2\sum_{i=1}^{\infty}\int_{(i-0.5)q}^{(i+0.5)q} p(x)|x - iq| dx$$

$$= \frac{1}{\eta}[1 + \frac{e^{-\eta q}}{1 - e^{-\eta q}}(2 - e^{-0.5\eta q} - e^{0.5\eta q}) - e^{-0.5\eta q}], \tag{16}$$

and

$$\sigma^2 = \int_{-\infty}^{+\infty} p(x)x^2 dx = \int_{-\infty}^{+\infty}\frac{\eta}{2}x^2 e^{-\eta|x|}dx = \frac{2}{\eta^2}. \tag{17}$$

thus based on (16) and (17)

$$D_i = \sum_{i=1}^{384} D(q) = \frac{384\psi_i}{\eta(2 - \psi_i)} = \frac{384 - \rho_i}{\sqrt{2}(384 + \rho_i)}\sigma_i. \tag{18}$$

By following similar steps as in (8)-(15), we get

$$\rho_i = \frac{\sqrt{w_i\sigma_i}}{\sum_{j=1}^{N}\sqrt{w_j\sigma_j}}\rho_{budget} + 384\frac{\sqrt{w_i\sigma_i} - \frac{1}{N}\sum_{j=1}^{N}\sqrt{w_j\sigma_j}}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{w_j\sigma_j}}, \tag{19}$$

and then simplify it to bit allocation model II:

$$\rho_i = \frac{\sqrt{w_i\sigma_i}}{\sum_{j=1}^{N}\sqrt{w_j\sigma_j}}\rho_{budget}. \tag{20}$$

We compare the two bit allocation models with the optimal solution by Lagrangian relaxation and the result is shown in Fig. 2. The perceptual PSNR is defined as

-$10\log_{10} D_{Frame}$. The result indicates that both models perform as closely as the optimal solution.

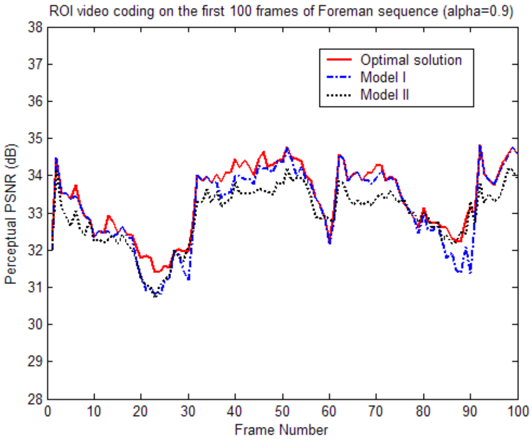


Figure 2 Comparison of weighted bit allocation models with optimal solution

Given a bit budget for a frame and using Eq. (15) or (20), we can optimally allocate the bits over the macroblocks within the frame to minimize the perceptual distortion defined in Eq. (1). We will use the bit allocation model II in the proposed system due to its simplicity.

3. ADAPTIVE BACKGROUND SKIPPING

At very low bitrate case, the Non-ROI regions are normally coarsely coded which results in low visual quality. On the other hand, in most cases of video telephony applications where background are Non-ROI, there are very limited movements in the background. Therefore, background skipping is a potential solution for reallocating bits to improve the quality of foreground and coded background regions as long as the skipping does not severely hurt the video fidelity.

The difference between background skipping and frame skipping is that the ROI for each frame is coded in the background skipping approach to ensure the good quality of ROI. Frame skipping is very helpful in plenty of applications, however in ROI video coding, it takes the risk of missing important information such as facial expressions; especially when α is set at a large value in Eq. (1), any distortion of ROI will be heavily punished and could degrade the overall performance. Therefore, background skipping is a better choice and it can generally save enough bits for improving ROI quality because the number of background macroblocks is dominant in typical video frames.

In this paper, we consider a prototype system in which we group every pair of frames into a unit. In each unit, the first background is coded while the second background is skipped (using predicted macroblocks with zero motion vectors) as shown in Fig. 3. In frame-level bit allocation, we assume that the content complexity of the video frames in a sequence is uniformly distributed and thus the bits are allocated uniformly among units. Within the unit, Eq. (20) is used for the bit allocation among macroblocks.

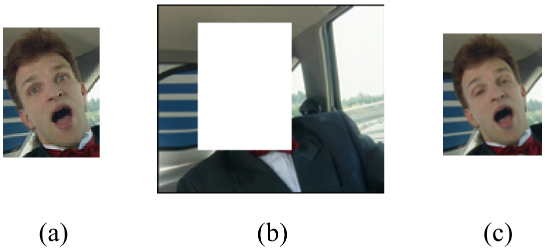


Figure 3 An example of a coded unit in the prototype system (a. ROI in frame 0; b. Non-ROI in frame 0; c. ROI in frame 1)

In the proposed system, background skipping in a unit is adaptively controlled based on the distortion caused by the skipping (D_{NonROI_skip}). For video sequences whose background contains large amount of motion, the skipping of important background information might severely undermine the system performance. Figure 4 shows the D_{NonROI_skip} statistics of the Carphone sequence.

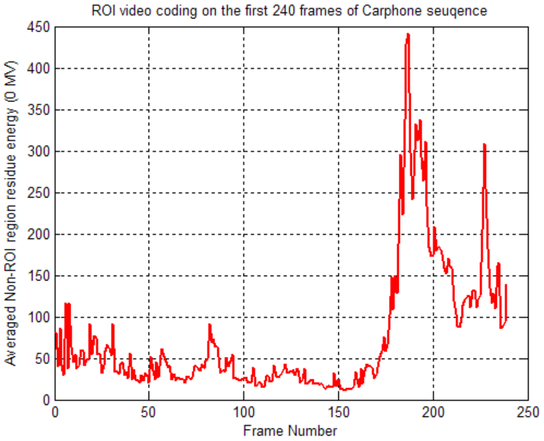


Figure 4 Statistics of distortion caused by background skipping

We use a distortion threshold to determine the background skipping mode. We define the threshold to be related to α and the statistics of the skipping distortion of the latest processed units. Let us denote by $\bar{D}_n$ the mean distortion of the latest n units, we use $\frac{\bar{D}_n}{2(1-\alpha)}$ as the threshold. We describe the adaptive background skipping algorithm as follows.

Step 0: Initialization (set $\bar{D}_n = 0$ and skipping mode=ON).

Step 1: Allocate ρ budget for the current (ith) unit by $\rho_{unit\ i} = \dfrac{\rho_{Segment} - \rho_{used}}{\frac{M}{2} - i}$, where M is the number of frames in the rate control segment, $\rho_{segment}$ is the number of ρ allocated to the segment, and ρ_{used} is the number of used ρ up to the current unit within the segment.

Step 2: Within the current unit, allocate bits for each macroblocks by Eq. (20). If the skipping mode is on, then no bits are assigned for the Non-ROI of the second frame.

Step 3: After the distortion for current unit is obtained, update $\bar{D}_n$ by $\bar{D}_n = (1-\eta)\bar{D}_{n-1} + \eta D_n$, where η is the learning factor and it is in the range of [0, 1].

Step 4: Updating the ρ statistics and get data for the next unit; if this is the last unit, go to step 6.

Step 5: Calculate D_{NonROI_skip} for the new unit, if $D_{NonROI\_skip} > \dfrac{\bar{D}_n}{2(1-\alpha)}$ then turn off the skipping mode. Go to step 1.

Step 6: Terminate the algorithm.

We encode the 180th to 209th frames of the Carphone sequence with three skipping modes (on, off and adaptively controlled) and show the results in Fig. 5. As expected the advantage of background skipping diminishes with the decreasing of α, and it is even more favorable for the mechanism without background skipping when α=0.5. The results also indicate that the behavior of the adaptive approach is always very close to the best solution for various α. In this experiment, we set η=0.25.

4. EXPERIMENTAL RESULTS

We conducted the simulations using the H.263 Profile 3 codec, and we tested Carphone and Foreman QCIF sequences at bitrates from 32kbps to 64kbps. In the experiments, we compared four different rate control approaches: (1) weighted bit allocation approach, where the bit allocation within a frame follows the model described by Eq. (20); (2) greedy algorithm, where the macroblocks are equally treated in the bit allocation as in [13]; (3) frame skipping algorithm, where the sequences are divided into units similarly as in section 3, and in each unit the second frame is skipped; (4) the proposed approach, which combines adaptive background skipping and optimized weighted bit allocation as described in sections 2 and 3.

The first experiment was conducted on Carphone sequence, and the results are shown in Figs. 6 and 7. As shown in Fig. 6, the proposed approach outperformed all other approaches in the whole bitrate range and the gain is up to 2dB. As an example, Fig. 7 shows the reconstructed frame for the greedy algorithm and the proposed approach. It is very clear that the result obtained from the proposed approach has a much better subjective visual quality compared to the other.

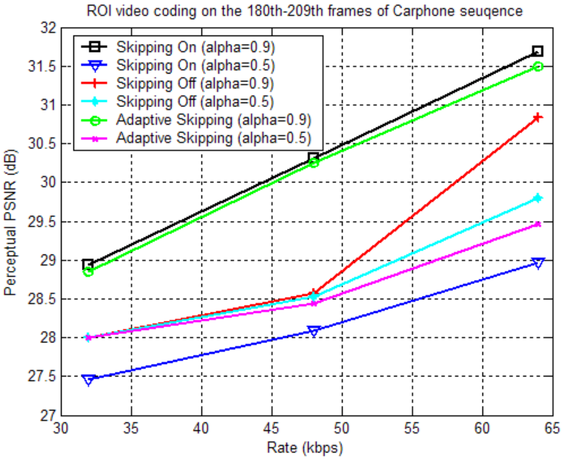


Figure 5 Comparison of adaptive skipping approach with other approaches

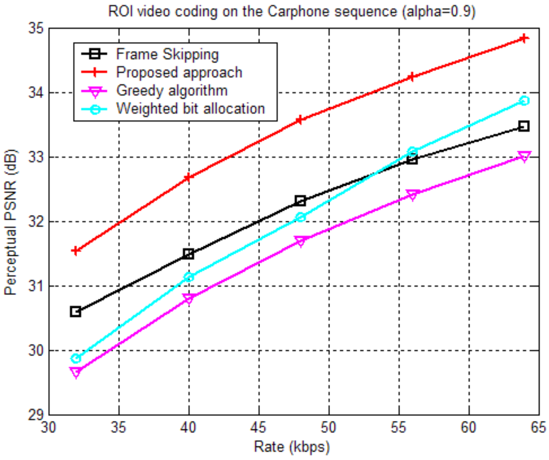
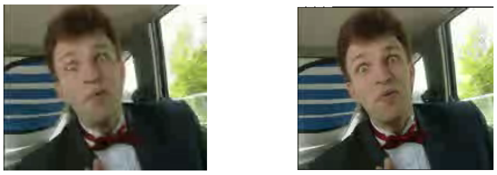


Figure 6 Comparison of various bit allocation approaches



(a) Greedy algorithm (b) Proposed approach

Figure 7 Comparison of various approaches at 40kbps

The second experiment was conducted on Forman sequence, and the results are shown in Figs. 8 and 9. Clearly, similar results were obtained as the first experiment. In Fig. 8, the frame skipping approach does not perform as well as in the first experiment, because the face of the Foreman sequence contains much larger motion compared to Carphone sequence, which means the frame skipping approach missed a lot of ROI information in this sequence and thus resulted in an unsatisfactory performance. The reconstructed video frames shown in Fig. 9 demonstrate the advantages of using the proposed approach.

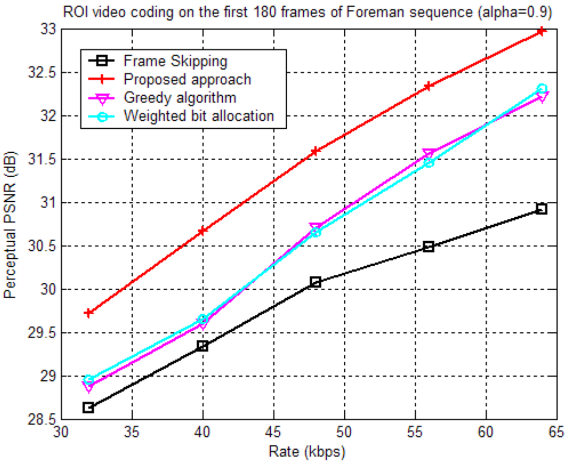
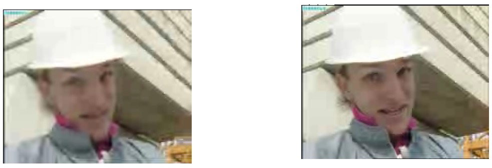


Figure 8 Comparison of various bit allocation approaches



(a) Greedy algorithm (b) Proposed approach

Figure 9 Comparison of various approaches at 40kbps

5. CONCLUSIONS

In this paper, we presented a novel ROI video coding algorithm for very low bitrate video applications such as wireless video telephony. First, we proposed two optimized weighted bit allocation schemes in ρ-domain for ROI video coding. Then, an adaptive background skipping approach was proposed which can work jointly with the weighted bit allocation models to achieve better performance. Experimental results indicate that the proposed algorithm has significant gains of up to 2dB over the other approaches.

REFERENCES

[1] A. Eleftheriadis and A. Jacquin, "Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates", *Signal Processing: Image Communications,* Vol. 7, No. 4-6, pp. 231-248, Nov. 1995.

[2] J. Lee, and A. Eleftheriadis, "Spatial-temporal model assisted compatible coding for low and very low bitrate video telephony", in *Proc. IEEE Int. Conf. Image Proc.*, Vol. 2, pp. 429-432, Lausanne, Switzerland, Sept. 1996.

[3] S. Daly, K. Matthews, and J. Ribas-Corbera, "Face-based visually-optimized image sequence coding", *ICIP'98*. Vol III, pp. 443-447.

[4] S. Daly, K. Matthews, and J. Ribas-Corbera, "As plain as the noise on your face: adaptive video compression using face detection and visual eccentricity models", *Journal of Electronic Imaging*, 10(1), Jan. 2001, pp. 30-46.

[5] D. Chai, and K. N. Ngan, "Face segmentation using skin-color map in videophone applications", *IEEE Trans. Circuits Systems for Video Technology*, Vol. 9, No. 4, June 1999, pp. 551-564.

[6] T. Adiono, T. Isshiki, K. Ito, T. Ohtsuka, D. Li, C. Honsawek and H. Kunieda, "Face focus coding under H.263+ video coding standard", in *Proc. IEEE Asia-Pacific Conf. Circuits and Systems*, Dec. 2000, Tianjin, China, pp. 461-464.

[7] M. Chen, M. Chi, C. Hsu and J. Chen, "ROI video coding based on H.263+ with robust skin-color detection technique", *IEEE Trans. Consumer Electronics*, Vol. 49, No. 3, Aug. 2003. pp. 724-730.

[8] C. Lin, Y. Chang and Y. Chen, "A low-complexity face-assisted coding scheme for low bit-rate video telephony", *IEICE Trans. Inf. & Syst.,* Vol. E86-D, No. 1, Jan. 2003. pp. 101-108.

[9] S. Sengupta, S. K. gupta, and J. M. Hannah, "Perceptually motivated bit-allocation for H.264 encoded video sequences", *ICIP'03*, Vol. III, pp. 797-800.

[10] 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", *ISCAS'2004*, Vol. III, pp. 805-808.

[11] D. Tancharoen, H. Kortrakulkij, S. Khemachai, S. Aramvith, and S. Jitapunkul, "Automatic face color segmentation based rate control for low bit-rate video coding", in *Proc. 2003 International Symposium on Circuits and Systems (ISCAS'03),* Vol. II, pp. 384-387.

[12] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications", *IEEE Trans. Circuits Systems for Video Technology*, Vol. 9, No. 1, pp. 172-185, Feb. 1999.

[13] Z. He and S. K. Mitra, "A linear source model and a unified rate control algorithm for DCT video coding", *IEEE Trans. Circuits and System for Video Technology*, Vol. 12, No. 11, Nov. 2002. pp. 970-982.

[14] H. Wang, G. M. Schuster, A. K. Katsaggelos, "Rate-distortion optimal bit allocation scheme for object-based video coding", *IEEE Trans. Circuits and System for Video Technology*, July-September, 2005. (to appear)

[15] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images", *IEEE Trans. Image Processing*, Vol. 9, No. 10, Oct. 2000. pp. 1661-1666.