PAPER

A Low-Complexity Face-Assisted Coding Scheme for Low Bit-Rate Video Telephony

Chia-Wen LIN[†], **Yao-Jen CHANG**[††], *and* **Yung-Chang CHEN**[††], *Nonmembers*

SUMMARY This paper presents a novel and practical face-assisted video coding scheme to enhance the visual quality of the face region in videophone applications. A low-complexity skin-color-based face detection and tracking scheme is proposed to locate the human face regions in realtime. After classifying the macroblocks (MBs) into the face and non-face regions, we present a dynamic distortion-weighting adjustment (DDWA) scheme to skip encoding the static non-face MBs, and the saved bits are used to compensate the face region by increasing the distortion weighting of the face MBs. The quality of the face regions will thus be largely enhanced. Moreover, the computation originally required for encoding the skipped MBs can also be saved. The experimental results show that the proposed method can significantly improve the PSNR and the subjective quality of face regions, while the degradation introduced on the non-face areas is relatively invisible to human perception. The proposed algorithm is fully compatible with the H.263 standard, and the low complexity feature makes it well suited to be implemented for real-time applications.

key words: *video coding, face detection, face tracking, low bit-rate video coding, rate control*

1. Introduction

With the rapid advance of video technology, digital video applications have become more and more popular in our daily life. Recently, H.261 and H.263 video coding standards were established for two-way video telephony applications such as videophone and video conferencing. Two-way video telephony applications require low-delay and real-time processing. For low bit-rate applications, the available bandwidth for video transmission is often less than 64 kb/s. The rate-control scheme, which decides the quantization step-size and monitors buffer fullness, plays an important role in the video encoder since it has a great impact on the video quality. Essential considerations in a rate-control scheme include the characteristics of video contents, channel characteristics, and buffer delay so as to optimally allocate the resource to maintain acceptable visual quality over low-bandwidth channels. In very low-bit-rate applications, the bit-allocation problem becomes critical in video quality.

Manuscript received March 29, 2002.
Manuscript revised June 17, 2002.

[†]The author is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan.

[††]The authors are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan.

Region-based rate-control strategies involving segmentation and dynamic bit-allocation of the areas of interest based on some quality criteria have been attracting many researchers' attention and considered as the promising approaches to very low-bit-rate video communications. In video telephony applications, the face area is usually the region of interest attracting the viewer's attention. It is thus worthwhile to allocate more bits to the face region to obtain a sharper face image by sacrificing the quality of the other regions to some acceptable extent [2]–[9]. For example, a 3-D subband coder which adopts face-model-assisted dynamic bit-allocation and pixel-based motion-compensation was proposed in [2] for encoding. Lee and Eleftheriadis [4] presented a method which encodes eyes, mouth, the remaining face area and background with different quantization parameters to achieve better perceived quality and effectively higher frame rate on face region. The selective encoding of various areas of interest without adapting to the real motions of the video contents, however, may result in annoying mismatch artifacts. Daley et al. [5] presented a quantization scheme of facial area using visual sensitivity which is a function of eccentricity in visual angle where the center of gaze is the reference angle of zero degree. The method is, however, not standard-compatible, because the face-location information needs to be sent to the decoder side, while the current standards do not provide the channel to transmit such overhead. More recently, Chai and Ngan [8] proposed a skin-color-based scheme for face segmentation, and then used two different quantizers, a finer quantizer and a coarser quantizer, to encode the face and non-face MBs, respectively. In their method, simply using two different quantization step-sizes may introduce significant distortions on non-face regions with high motion activities and result in inaccurate bit-rate control.

The low-delay and limited computing power (e.g., for mobile videophone and hand-held devices, etc.) constraints in video telephony applications would prevent one from using sophisticated methods for face detection and tracking. A general face-detection scheme may need to tackle multiple-face situations and segment out the accurate face contours, which will involve sophisticated segmentation procedures, thereby increasing the implementation cost. In videophone applications, there will be typically only one face appearing on the dis-

play. Moreover, for face-assisted video coding in current block-based video coding standards, the face region need not be segmented very accurately, since the coding granularity is at the block level rather than the pixel level. Taking these points into account, the implementation cost can be reduced largely. In addition, a standard-compatible scheme without the need of sending non-standard overhead is desirable.

In this paper, we present a low-complexity skin-color-based approach to real-time classifying the pixels into the face and non-face classes. After pixel classification, a double integral projection method is subsequently used to segment out the face area. Based on the H.263 TMN8 rate-control framework [10], [11], we propose a DDWA scheme which can enhance the quality of the face regions significantly by dropping the static non-face MBs. Since most of the degraded MBs are not the region of interest, the introduced distortion is thus nearly invisible to the viewer's perception. Another advantage of the proposed DDWA method is that the computation (DCT, quantization, inverse quantization, and IDCT) required for encoding the skipped MBs can be saved, thereby achieving further computation reduction.

The rest of this paper is organized as follows. In Sect. 2, a skin-color-based face detection and tracking scheme is presented. The proposed dynamic distortion-weighting adjustment scheme for face region enhancement is described in Sect. 3. Section 4 shows the experimental results of the proposed algorithms and the comparison with the H.263 TMN8 method. Finally, conclusions are drawn in Sect. 5.

2. Real-Time Skin-Color Based Face Detection and Tracking

A skin-color-based approach is proposed for fast face detection and tracking. As depicted in Fig. 1, skin-color classification is performed on each pixel according to its skin-color probability, resulting in a binary map indicating skin-color pixels. An efficient double integral projection algorithm is proposed to detect the face region on the binary map. After determining the face region of the first frame, the skin-color model is adapted to the user, and the face-tracking mode is activated to fast locate the face block in the following sequences.

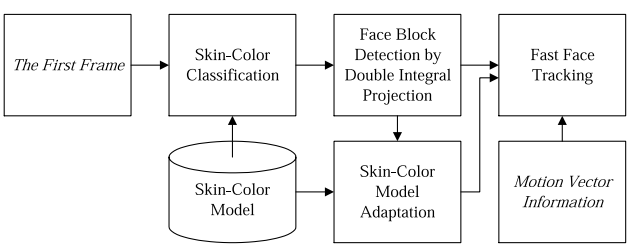


Fig. 1 The proposed face detection and tracking algorithm.

The motion vector information available from the video encoder is used to estimate the center of the face region for face tracking. The motion information can also be used to assist removing the background color noises, since the background areas are often of less motion activity.

2.1 Skin-Color Classification with a Gaussian Probability Model

As explored in the literature [12]–[14], skin-color of human beings in different races forms a very compact area in the chrominance space such as C_b-C_r space [12] and normalized r-g space [13], [14]. Therefore, a Gaussian model $N(\mathbf{m}, \mathbf{\Sigma})$ [13] can be utilized for representing the skin-color distribution parameterized by its mean $\mathbf{m} = (\bar{c}_b, \bar{c}_r)$ and covariance $\mathbf{\Sigma}$. Thus, each pixel with color components $\mathbf{x} = (c_b, c_r)$ can be classified as the skin-color class if its p.d.f. value is larger than a threshold. That is

$$\mathbf{x} \in \text{skin-color class if } p(\mathbf{x}) > TH_{\text{skin}}$$
$$\text{where } p(\mathbf{x}) = \frac{1}{(2\pi)|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right) \tag{1}$$

The Gaussian probability model of skin-color can be built using a number of test videos in advance. Furthermore, the adaptation method described in [15] can be utilized to refine the Gaussian model to adapt to the current statistics. The refined model will tend to be closer to the skin-color distribution of the current user so as to filter out the background noise more easily.

2.2 Face Detection and Tracking

After skin-color classification, a binary image representing the skin-color pixels is generated. The binary image is subsequently median filtered to remove the granular noises due to possible false classifications. In contrast to the binary template matching algorithm proposed in [12], a "double integral projection" scheme is proposed for face block detection and tracking on the binary image. The proposed scheme not only reliably detects the face area with a very fast speed but also handles the possible situation when the user is dressed in skin-colored clothes. The process of the proposed double integral projection is as the following steps:

Step 1. Project the median-filtered binary image $F(x, y)$ in Fig. 2(b) using the horizontal integral projection as shown in Fig. 2(c):

$$H(y) = \sum_x F(x, y). \tag{2}$$

Step 2. Convert the projected value $H(y)$ to a binary image $H'(x, y)$, and perform the vertical integral

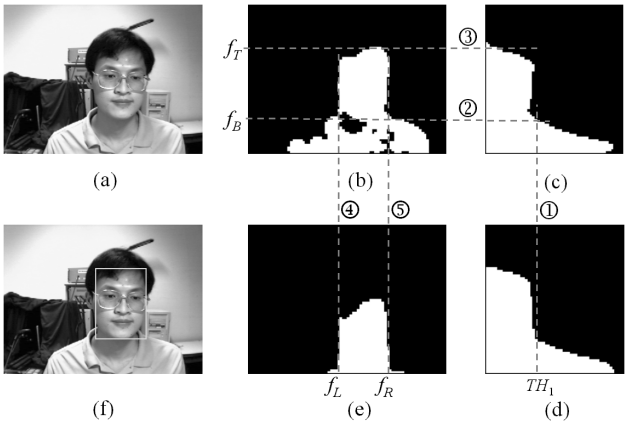


Fig. 2 The proposed double integral projection process for face detection and tracking. (a) The original image; (b) the median-filtered binary image generated from skin-color classification; (c) the horizontal integral projection image projected from (b); (d) the vertical integral projection image projected from (c); and (e) the vertical integral projection image projected from (b) between f_T and f_B; (f) the detected face block overlapped on the original image. (Note: The index number represents the generating order of each line.)

projection as shown in Fig. 2(d):

$$H'(x,y) = \begin{cases} 1, & \text{if } H(y) \geq x \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$V(x) = \sum_{y} H'(x,y). \tag{4}$$

Step 3. Determine the threshold TH_1 using second-order derivatives of $V(x)$:

$$TH_1 = \arg\max_{x} \left\{ \frac{d}{dx^2} V(x) \right\}. \tag{5}$$

Step 4. Search from the bottom of the horizontally projected image and choose the lowest position whose value is under TH_2 as the bottom of the face block, f_B,

$$f_B = \max\{y | TH_2 < H(y) < TH_1\}$$
$$\text{with } TH_2 = k \max(H(y)).$$

The setting of lower bound TH_2 is utilized to handle the case when the user is not dressed in skin-colored clothes. And k is a small number which is set to be 0.1 in our experiment.

Step 5. The top of the face block, f_T, can be found by simply thresholding the horizontally projected image above f_B.

Step 6. The left and right boundaries, f_L and f_R, of the face block can be easily determined from the vertically projected image taken from the portion of the binary image between f_T and f_B as shown in Fig. 2(e).

Note in Step 2, the vertical integral projection is performed on the result of the horizontal integral projection formed in Step 1. Therefore good threshold value can be derived for segmentation of the face region and the shoulder region even when the user is dressed in skin-colored clothes as illustrated in Fig. 2(f). In case of background with large skin-color regions, the luminance regularization proposed in [8] can be utilized for filtering out homogeneous regions.

The TH_{skin} in Eq. (1) can be determined with an automatic thresholding scheme when processing the first frame. Experimental results indicate that a range of threshold values leading to reliable segmentation results usually has the following properties: (1) small face block-size variation, (2) medium to high fullness of skin-like pixels within the detected face block, and (3) small face center location variation, with respect to the threshold change. Thus, the value of TH_{skin} can be chosen as the median of a threshold range possessing these properties.

After the face detection process, a face-tracking mode is activated to fast locate the face block in the following frames. In the face-tracking mode, the motion information (e.g., average motion vector) of the detected face block in the previous frame is used to determine the possible center of the face block in the current frame. Instead of processing the whole image, color classification is performed within a reduced area determined by enlarging the previously detected face block translated by the initial motion vector. The frame differences are used to detect some exceptional cases (e.g., content change) to determine whether a face detection process should be re-initialized to avoid possible false detection with the tracking mode. The pixel differences between two consecutive frames can also be used to remove the skin-color-like noise belonging to a still background since the face region is usually a moving object leading to lager pixel differences, while the differences of the still background pixels are usually much smaller.

2.3 Complexity Analysis

The calculation of the proposed algorithm has to perform skin-color classification and face detection and tracking. For the skin-color classification, we calculate the p.d.f. values of each pixel from the downsampled C_b and C_r frames with Eq. (1). The covariance matrix remains constant after the adaptation procedure is performed. Therefore, the denominator and the exponential function can be combined to the TH_{skin} term:

$$x \in \text{skin-color class if } p'(x) < TH'_{\text{skin}}$$
$$\text{where } p'(x) = (x-m)^T \Sigma^{-1} (x-m)$$
$$\text{and } TH'_{\text{skin}} = -2 \ln \left((2\pi) |\Sigma|^{\frac{1}{2}} \times TH_{\text{skin}} \right).$$

Hence, the number of total calculations for a p.d.f. value computation is reduced to 6 multiplications and 5 additions. To further reduce the computations, a lookup table of $p'(x)$ values can also be constructed during run-time so that the $p'(x)$ values of most pixels

will be found in the lookup table in a short duration since the contents of a video-telephony image sequence is temporarily highly correlated.

For the face detection, the integral projection is performed three times for the down-sampled binary image. Therefore, less than $3w \times h/16$ additions are required for an image of size $w \times h$, which consumes very little calculations.

3. Proposed Face-Assisted Macroblock-Layer Bit Allocation

A model-based rate-distortion constrained rate control technique was proposed in [10], and was adopted in the ITU-T test model TMN8 [11] of H.263+ for low-delay, two-way, real-time applications. The following models are proposed in [10] to estimate the produced bits and the distortion of a MB:

$$B_i = A\left(K\frac{\sigma_i^2}{Q_i^2} + C\right), \tag{6}$$

$$D_i = \alpha_i^2 \frac{Q_i^2}{12}, \tag{7}$$

where K is the model parameter with a typical value of $e/\ln 2$ [10], A is the number of pixels in a MB, σ_i^2 is the MB variance of the motion-compensated difference frame, C is the average rate to encode the motion vectors and the bit-stream header for the frame, and parameter Q_i is the quantizaion step-size of the i-th MB, and α_i is the distortion weight.

The optimal quantization step-size can be decided by finding the solution of

$$Q_1^* \ldots, Q_N^* = \underset{Q_1,\ldots,Q_N}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} \alpha_i^2 \frac{Q_i^2}{12} \tag{8}$$

$$\text{subject to } \sum_{i=1}^{N} B_i \le B_{\text{target}}$$

where N is the number of macroblocks in a frame, and B_{target} is the target bit-count for the frame.

The Lagrange multiplier can be used to solve this constrained optimization problem. The optimized quantization step-size was derived in [10] as follows:

$$Q_i^* = \sqrt{\frac{AK}{(B_{\text{target}} - ANC)}\frac{\sigma_i}{\alpha_i}\sum_{j=1}^{N}\alpha_j\sigma_j}, i = 1, 2, \ldots, N \tag{9}$$

In Eq. (7), the parameter α_i^2 represents the distortion weight, which provides an efficient means to dynamically control the MB quality. With this distortion model, dynamic distortion-weighting adjustment of different regions of interest can be performed by assigning different weights to the MBs belonging to different regions. We propose to classify the MBs into three regions: face region, active non-face region, and static non-face region. The rule used in our proposed algorithm to classify the active and static non-face MBs is as follows:

```
if ( MB_i ∉ Face_Region)
    if (SAD_i < TH_SAD) and (SAMV_i < TH_MV)
        MB_i ∈ Static_Nonface_Region;
    else
        MB_i ∈ Active_Nonface_Region;
```

where SAD_i indicates the sum of absolute difference of the i-th MB, and $SAMV_i$ is the sum of the absolute motion vector of the i-th MB as follows:

$$SAD_i = \sum_{x,y \in MB_i} |f_n(x,y) - f_{n-1}(x + MV_{i,x}, y + MV_{i,y})| \tag{10}$$

and

$$SAMV_i = |MV_{i,x}| + |MV_{i,y}| \tag{11}$$

where $f_n(x,y)$ represents the pixel value at (x,y) position of frame n, $(MV_{i,x}, MV_{i,y})$ is the motion vector associated with the i-th MB. Since the values of SAD and motion vectors are the by-products of motion estimation, the extra computational cost required for the above classification rule is very low. Because the non-face MBs may belong to the background or the human body, larger distortions in the non-face regions are tolerable to viewer's perception since the face region is the focus in video telephony applications. We propose to skip encoding the static non-face MBs by setting its associated coded macroblock indication (COD) bit to 1 in the H.263 syntax, and the saved bits are used to compensate the quality of the face and active non-face MBs using larger weights. The proposed MB-layer DDWA scheme is summarized as follows:

```
for i = 1 to N
{
Set the initial weighting values as follows:
    α_ini = { 2(B_target/AN)(1 − σ_i) + σ_i, if B_target/AN < 0.5
            { 1                            , otherwise
if (MB_i ∈ Face_Region)
    α_i = α_ini × K_face
else if (MB_i ∈ Static_Nonface_Region)
    α_i = 0
else
    α_i = α_ini
endif
}
```

for $i = 1$ to N
{
calculate

$$Q_i = \min\left(31, \sqrt{\frac{AK}{(B_{\text{target}} - ANC)} \frac{\sigma_i}{\alpha_i} \sum_{j=1}^{N} \alpha_j \sigma_j}\right)$$

}

In the above algorithm, the initial distortion weight, α_{ini}, is set according to the suggestion in [10]. After face segmentation, the distortion weight, α_i, for face MBs is magnified by a ratio K_{face}. The factor K_{face} represents a relative importance weight of the face region, and the H.263+ TMN8 rate control will dynamically allocate more bits to the face MBs accordingly. Since K_{face} is a relative ratio with respect to other regions, and most faces present similar texture, it is reasonably good to keep K_{face} constant regardless of the face size detected. With the dynamic weighting adjustment, the quality of the face region can thus be effectively enhanced. On the other hand, the quality of the static non-face region will be sacrificed, and only a small difference appears on the active non-face region. The improvement on the face region is significant, while the degradation on the non-face region due to DDWA is relatively invisible to human perception in video telephony applications as will be shown from the experimental results. Moreover, the computations (DCT, quantization, inverse quantization, and IDCT) required for encoding the skipped MBs can be saved, thus the computational cost can be further reduced.

4. Experimental Results

In our experiments, the frame-layer bit-allocation scheme proposed in TMN8 is employed, and the implementation is based on the UBC H.263+ public domain software [16]. Four QCIF (176×144) test sequences: "Miss_am," "Suzie," "Foreman" and "Carphone" are used to demonstrate the performance of the proposed algorithm, where "Miss_am" and "Suzie" sequences are encoded at 36 kb/s, and "Foreman" and "Carphone" are encoded at 64 kb/s. The sampling rate and the encoding frame-rate of these four sequences are all 30 frames per second. Table 1 lists the typical coding parameters and threshold values used in the experiments.

Table 2 compares the PSNR quality of the proposed DDWA and the TMN8 methods. The simulation results show that the DDWA method can significantly improve the PSNR quality of the face regions. In our experiments, the PSNR improvement ranges from 0.4 dB to 2 dB, depending on the complexity levels of the video contents. In general, if most of the background area is uniform and motionless, the gain on face region enhancement will be less, since the number of saved bits from skipping the static non-face MBs will be relatively small. This is why the PSNR improvement on the face regions of the "Suzied" sequence is the least among the four sequences.

Table 1 Typical values of parameters and thresholds used in the experiments.

Parameter / threshold	Value
TH_{SAD}	$\frac{2}{N}\sum_{i=1}^{N} SAD_i$
TH_{MV}	1
K	$e/\ln 2$
K_{face}	16
TH_{face}	automatically determined

Table 2 Average PSNR comparison of the proposed DDWA and the TMN8 rate control schemes. (unit: dB)

	PSNR of Frame		PSNR of Face Region	
	TMN8	DDWA	TMN8	DDWA
Miss_am	37.32	36.86	32.56	33.59
Suzie	31.75	31.45	29.85	30.25
Foreman	29.48	28.65	29.74	31.21
Carphone	30.42	29.36	29.97	31.92

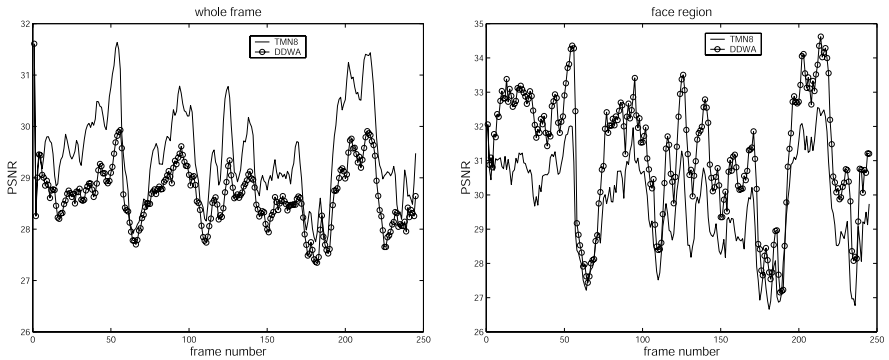
Table 3 Maximum PSNR improvement and face degradation on the whole frame and face region with the proposed DDWA and the TMN8 rate control scheme. (unit: dB)

	PSNR of Frame		PSNR of Face Region	
	Max. PSNR Improvement	Max. PSNR Degradation	Max. PSNR Improvement	Max. PSNR Degradation
Miss_am	0.28	1.16	2.15	0.00
Suzie	0.25	1.19	1.66	0.59
Foreman	0.11	1.85	3.55	0.81
Carphone	0.58	2.96	4.80	0.51

Figures 3 and 4 show the performance comparisons between the proposed methods and the TMN8 rate control for the "foreman" and "carphone" sequences respectively. The proposed method can achieve significant quality enhancement on face regions for all the test sequences most of the time. Table 3 lists the maximum performance improvement and degradation on the whole frame and face region of the four test sequences. The maximally enhanced and degraded frames are also compared in Figs. 3 and 4. It is evident that the face region improvement is highly visible to human perception, while the distortions on the background area of the maximally degraded frames are relatively invisible since we seldom pay attention to the backgrounds in video telephony applications.

5. Conclusions

In this paper, we have proposed a standard-compatible video coding scheme which can effectively enhance the visual quality of face regions with low extra complexity. The novelty of our proposed algorithm includes the following. First, we have presented a low-complexity skin-color-based face detection and tracking scheme. After skin-color classification, a double integral projec-



(a) (b)

(c) (d)

(e) (f)

Fig. 3 (a), (b) PSNR performance comparison of the proposed method with the TMN8 rate control on "Foreman" sequence at 64 Kb/s; (c) frame #57 coded using TMN-8 (face: 28.90 dB); (d) frame #57 coded using the proposed method (face: 32.35 dB) (maximally improved frame on the face region); (e) frame #201 coded using TMN8 (whole frame: 30.92 dB); (f) frame #201 coded using the proposed method (whole frame: 29.07 dB) (maximally degraded frame on the whole frame).

tion method is subsequently used to segment out the face MBs from the classified skin-color pixels. The proposed face detection and tracking algorithm can process more than 30 CIF frames per second when implemented on a PC Pentium II 300 machine. Secondly, we have proposed a dynamic distortion weighting adjustment method by skipping the encoding of the static non-face MBs to enhance the quality of face MBs and reduce the computational cost. The experimental results show that the proposed algorithm can effectively enhance the

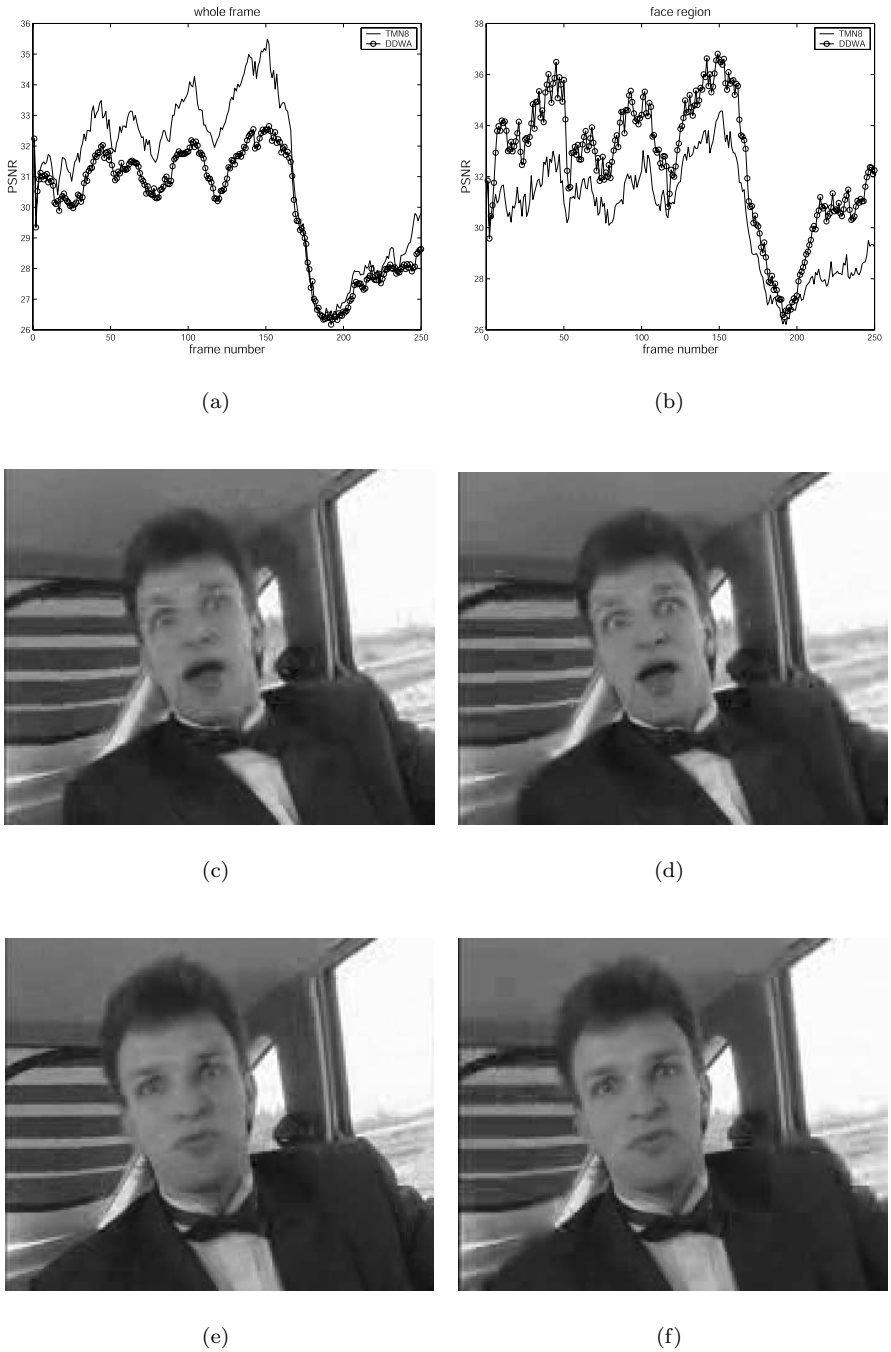


Fig. 4 (a), (b) PSNR performance comparison of the proposed method with the TMN8 rate control on "Carphone" sequence at 64 Kb/s; (c) frame #50 coded using TMN-8 (face: 30.98 dB); (d) frame #50 coded using the proposed method (face: 35.78 dB) (maximally improved frame on the face region); (e) frame #151 coded using TMN8 (whole frame: 35.49 dB); (f) frame #151 coded using the proposed method (whole frame: 32.53 dB) (maximally degraded frame on the whole frame).

visual quality of face regions at the cost of introducing some degradation on the inactive non-face regions. The performance improvement on face region ranges from 0.4 to 2 dB for the four test sequences. The degradation is, however, relatively invisible to human perception in video telephony applications. The extra computational complexity of the proposed algorithm is quite low thus making it well suited for real-time applications. Furthermore, the proposed algorithm does not change the syntax of the current video coding standards, thus can

be integrated into the standard-compliant commercial products.

References

[1] ITU-T Draft Recommendation H.263, "Video coding for low bit-rate communication," May 1997.
[2] 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 Communication, vol.7, no.4-6, pp.231–248, Nov. 1995.
[3] Hartung, A. Jacquin, J. Pawlyk, J. Rosenberg, H. Okada, and P. Crouch, "Object-oriented H.263 compatible video coding platform for conferencing applications," IEEE J. Sel. Areas Commun., vol.16, no.1, pp.42–55, Jan. 1998.
[4] J.-B. Lee and A. Eleftheriadis, "Spatio-temporal model assisted compatible coding for low and very low bitrate video telephony," Proc. IEEE Int. Conf. Image Proc., vol.2, pp.429–432, Lausanne, Switzerland, Sept. 1996.
[5] S. Daley, K. Matthews, and J. Ribas-Corbera, "Face-based visually-optimized image sequence coding," Proc. IEEE Int. Conf. Image Processing, pp.443–447, Chicago, IL, USA, Oct. 1998
[6] K. Ishikawa and O. Nakamura, "Very low bit-rate coding based on a method of facial area specification," IEEE Canadian Conference on Electrical and Computer Engineering, vol.1, pp.269–272, Waterloo, Ontario, Canada, May 1998.
[7] L. Ding and K. Takaya, "H.263 based facial image compression for low bitrate communications," Proc. IEEE Conf. on Comm., Power, and Computing (WESCANEX'97), pp.30–34, Winnipeg, MB, USA. May 1997.
[8] D. Chai and K.N. Ngan, "Face segmentation using skin-color map in videophone applications," IEEE Trans. Circuits Syst. Video Technol., vol.9, no.4, pp.551–564, June 1999.
[9] S. Lee, M.S. Pattichis, and A.C. Bovic, "Rate control for foveated MPEG/H.263 video," Proc. IEEE Int. Conf. Image Processing, Chicago, pp.365–369, Chicago, IL, USA, Oct. 1998.
[10] J. Ribas-Corbera and S.-M. 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.
[11] ITU-T/SG16, "Video codec test model, TMN8," Portland, June 1997.
[12] 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, Aug. 1997.
[13] J. Yang and A. Waibel, "A real-time face tracker," Proc. 3rd IEEE Workshop on Applications of Computer Vision, pp.142–147, Sarasota, FL, USA, Dec. 1996.
[14] N. Oliver, A. Pentland, and F. Berard, "LAFTER: A real-time lips and face tracker with facial expression recognition," Pattern Recognition, vol.33, no.8, pp.1369–1382, Aug. 2000.
[15] C.E. Priebe, "Adaptive mixtures," Journal of the American Statistical Association, vol.89, no.427, pp.796–806, 1994.
[16] Image Processing Lab, University of British Columbia, "H.263+ encoder/decoder," TMN (H.263) codec, Feb. 1998.
[17] C.-W. Lin, Y.-J. Chang, and Y.-C. Chen, "Low-complexity face-assisted video coding," Proc. IEEE Int. Conf. Image Processing, pp.207–210, Vancouver, BC, Canada, Sept. 2000.

Chia-Wen Lin received the M.S. and Ph. D. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1992 and 2000, respectively. Dr. Lin joined the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, in August 2000, where he is currently an Assistant Professor. Before that, he was a Section Manager of the CPE and Access Technologies Department at the Computer and Communications Research Laboratories, Industrial Technology Research Institute (CCL/ITRI), Taiwan. During April to August 2000, he was with the Information Processing Lab, Department of Electrical Engineering, University of Washington, as a Visiting Scholar. His research interests include video coding and networked multimedia technologies.

Yao-Jen Chang received his B.S. and Ph. D. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1996 and 2002, respectively. His research interests include computer vision, computer graphics, and multimedia signal processing.

Yung-Chang Chen received his B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1968 and 1970, respectively, and the Ph. D. (Dr. Ing.) degree from Technische Universitaet Berlin, Germany, in 1978. He joined the Department of Electrical Engineering at National Tsing Hua University, Hsinchu, Taiwan in 1978. He was Chair of Department of Electrical Engineering at National Central University, Chungli, Taiwan from 1980 to 1983, and Chair of Department of Electrical Engineering at National Tsing Hua University from 1992 to 1994. He is now a Professor of the Department of Electrical Engineering in National Tsing Hua University, Hsinchu, Taiwan. He is now Senior Member of IEEE and serves as Chair of CES, Taipei Chapter. His current research interests include multimedia signal processing, digital video processing, medical imaging, computer vision and pattern recognition.