# LOW-COMPLEXITY FACE-ASSISTED VIDEO CODING

Chia-Wen Lin
Dept. Computer Science and Information Eng.
National Chung Cheng University
Chia-Yi, Taiwan 621, R.O.C.
ljw@ieee.org

Yao-Jen Chang and Yung-Chang Chen
Dept. Electrical Eng.
National Tsing Hua University
Hsinchu, Taiwan 300, R.O.C.
{kc,ycchen}@benz.ee.nthu.edu.tw

## ABSTRACT

This paper presents a novel face-assisted video coding scheme to enhance the visual quality of the face regions in video telephony applications. A skin-color based face detection and tracking scheme is proposed to locate the face regions in real-time. After classifying the macroblocks into the face and non-face regions, we present a dynamic distortion weighting adjustment (DDWA) scheme to drop the static non-face macroblocks, and the saved bits are used to compensate the face region by adjusting the distortion weighting of the face macroblocks. The quality of face regions will thus be enhanced. Moreover, the computation originally required for the skipped macroblocks 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 insensitive to human perception. The proposed algorithm is fully compatible with H.263 standard, and the low complexity feature makes it well suited to implement for real-time applications.

## 1. INTRODUCTION

Two-way video telephony applications require low-delay and real-time processing. For low bit-rate applications, such as videophone over PTSN through 33.6 kb/s modem, the available bandwidth for video transmission is often less than 24 kb/s. Rate control scheme, which decides the quantization step-size and monitors buffer fullness, plays an important role in the video encoder which can greatly affect the video quality. To maintain acceptable visual quality and meet the real-time requirement over low-bandwidth channels, an efficient rate control scheme with low computing demand is required which is able to take into consideration the significance of video contents, channel statistics, and viewer's visual perception so as to optimally allocate the bit resource.

In video telephony applications, often the talker's head-and-shoulder image is viewed on the display. Thus the face area is often the region of interest attracting most of the viewer's attention. It's thus worthwhile to allocate more bits the face region to obtain sharper face quality by sacrificing the quality of the other regions to some acceptable extent. The ideas of allocating more bits to the regions of viewer's interest, particularly the face region, is not new [1-3]. Eleftheriadis and Jacquin described a face model-assisted coding method in [1] to selectively encode different areas of interest to obtain sharper face quality. The face region is detected using edge thresholding and ellipse model fitting. Dale *et al*. [2] 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 proposed in [2] is, however, incompatible with H.263 standard, because the region of interest information needs to be transmitted to the decoder side, and H.263 does not provide the channel to transmit such overhead. Chai and Ngan [3] proposed a skin-color based face detection approach and used the segmentation result for H.261 compatible video coding. macroblocks belonging to the segmented face region are classified as the foreground macroblocks and the others are the background macroblocks. Two different quantizers, a finer quantizer and a coarser quantizer, are used to encode the foreground and background macroblocks, respectively. With the foreground/background encoding method, MQUANT information should be sent as an overhead which will somehow increase the bit rate.

In this paper, we present a low-complexity skin-color based approach for real-time face detection and tracking in video telephony applications. We propose to use a Gaussian model to classify the pixels into skin-color and non-skin-color classes. We then develop a novel double integral projection method to fast segment out the face blocks. Based on the H.263 TMN8 rate control framework [5], we propose a dynamic distortion weighting adjustment (DDWA) scheme to enhance the quality of the face regions by dropping the static non-face macroblocks. Note that, due to the skipping of the inactive non-face macroblocks, this may introduce PSNR degradation in view of the whole frame. Since most of the degraded macroblocks are not the regions of interest, the introduced distortion is thus relatively insensitive to viewer's perception. Therefore, it is worthwhile to sacrifice the

quality of the regions of less importance and use the saved bits to enhance the quality of the highly focused regions, such as face regions in video telephony applications.

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

## 2. THE PROPOSED SKIN-COLOR BASED FACE DETECTION METHOD

Recently, some algorithms that utilize color information to detect the faces and facial features were presented. Because processing color information is much faster than other complicated methods, using skin-color for detecting the faces can have many advantages. However, the skin-color is different from person to person, and different video cameras and various lighting conditions may influence the color distribution. The problem is resolved by modeling the skin-color in a statistical approach [4]. Statistical experiments discovered that although skin-colors of different people appear to vary over a wide range, they differ less in chrominace than in brightness. A Gaussian model N($m$, $\Sigma^2$) is utilized in our method for representing the skin-color model with its mean $m = (\overline{c}_b, \overline{c}_r)$ and covariance $\Sigma$ calculated by (1)-(3)

$$\overline{c}_b = \frac{1}{N}\sum_{i=1}^{N} c_b^i \tag{1}$$

$$\overline{c}_r = \frac{1}{N}\sum_{i=1}^{N} c_r^i \tag{2}$$

$$\Sigma = \begin{bmatrix} \sigma_{c_b c_b} & \sigma_{c_b c_r} \\ \sigma_{c_r c_b} & \sigma_{c_r c_r} \end{bmatrix} \tag{3}$$

where $c_b^i$ and $c_r^i$ represent the values of the $C_b$ and $C_r$ components of the $i$th pixel. Each pixel in an image can be classified into skin-color class and non-skin-color class by calculating the its p.d.f. value from the Gaussian model. A pixel is classified as the skin-color class if the probability of the pixel belonging to the skin-color class is larger than a threshold. That is

$$x \in \text{skin-color class if } p(x) > TH_{skin}$$

$$\text{where } p(x) = \frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)\right) \tag{4}$$

After classifying each pixel in the incoming image, a binary image representing the skin-color pixels is generated. A "double integral projection" scheme is proposed for face block detection and tracking on the binary image. The process of the proposed double integral projection is as follows:

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

$$H(y) = \sum_x F(x, y) \tag{5}$$

**Step 2.** Convert the projected value $H(y)$ to a binary image $H'(x, y)$, and perform the vertical integral projection as shown in Fig. 1(d):

$$H'(x, y) = \begin{cases} 1, \text{if } H(y) \geq x \\ 0, \text{otherwise} \end{cases} \tag{6}$$

$$V(x) = \sum_y H'(x, y) \tag{7}$$

**Step 3.** Determine the threshold $TH_1$ with respect to the maximum value in Fig. 1(d), and the smallest position $p$ whose value is under the threshold is taken as the threshold value $TH_2$ for the horizontal integral projected image $H'(x, y)$,

$$TH_1 = k \max(V(x))$$

$$TH_2 = p = \min\{x \mid V(x) < TH_1\}$$

In our experiments, the value $k$ can be chosen from 1/3 to 1/2.

**Step 4.** Search from the bottom of the horizontal projected image and the lowest position whose value is under $TH_2$ is the bottom of the face block, $f_{bottom}$,

$$f_{bottom} = \max\{y \mid H(y) < TH_2\} \tag{8}$$

**Step 5.** The top of the face block, $f_{top}$, can be found by simply thresholding the horizontal projected image above $f_{bottom}$.

**Step 6.** The left and right boundaries, $f_{left}$ and $f_{right}$, of the face block can be easily determined from the vertical projected image taken from the portion of the binary image between $f_{top}$ and $f_{bottom}$ as shown in Fig. 1(e).

## 3. FACE-ASSISTED DYNAMIC RATE CONTROL IN VIDEO CODING

We propose to classify the macroblocks into three categories: face macroblocks, active non-face macroblocks, and static non-face macroblocks, then apply different distortion weighting factors to difference classes of macroblocks accordingly. The rule used in our proposed algorithm to classify the active and static non-face macroblocks is as follows:

*if* $(MB_i \notin Face\_Region)$

  *if* $(SAD_i < TH_{SAD})$ && $(Sum\_of\_MV_i < TH_{MV})$,

    $MB_i \in Static\_Non\_Face\_Region$;

  *else*

    $MB_i \in Active\_Non\_Face\_Region$;

where $SAD_i$ and $Sum\_of\_MV_i$ are defined as follows:

$$Sum\_of\_MV_i = \left|MV_{i,x}\right| + \left|MV_{i,y}\right|. \quad (9)$$

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

Because the non-face macroblocks may belong to background or human body, larger distortions in the non-face regions are tolerable to viewer's perception since the face region is usually the focus in video telephony applications. We propose to skip the static non-face macroblocks, and the saved bits are used to compensate the quality of the face and active non-face macroblocks using different weightings. The proposed macroblock layer DDWA approach is summarized as follows:

*for i = 1 to N*

*Set* $\alpha_i = \sigma_i$ *as the initial approximation*

*if* $(MB_i \in Face\_Region)$

$\alpha_i = K \cdot \sigma_i$

*else*

*if* $MB_i \in Static\_Non\_Face\_Region$

$\alpha_i = 0$

*endif*

*endif*

*calculate* $Q_i$ *using the method proposed in [5]*

$$Q_i^* = \sqrt{\frac{AK}{(B - ANC)} \frac{\sigma_i}{\alpha_i} \sum_{i=1}^{N} \alpha_i \sigma_i},$$

where *K* is the weight to determine the enhancement effect, N is the number of pixels in a macroblock, and A and C are the parameter models used in TMN8 rate control [5]. With the above algorithm, the distortion weighting factors, $\alpha_i^2$, will be magnified by a ratio for the face MB's, thus finer quantization parameters will be used and more bits will be allocated to the face macroblocks. The quality on face region can thus be effectively enhanced. The improvement on the face region is significant as will be shown from the experimental results, while the degradation on the non-face region due to DDWA is relatively invisible to human perception in video telephony applications. Moreover, another advantage of the proposed DDWA method is that the computations (DCT, quantization, inverse quantization, and IDCT) required for the skipped macroblocks can be saved, thus the computational cost can be further reduced.

## 4. EXPERIMENTAL RESULTS

In our experiment, the macroblock layer bit allocation scheme of TMN8 [5] is employed, and the implementation is based on the UBC H.263+ source code [6]. Four QCIF (176x144) 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 I shows the experimental results and indicates that the proposed algorithm can effectively enhance the visual quality of face regions at the cost of introducing some degradation on the non-face regions. The performance improvement on face region ranges from 0.4 to 2 dB for the four test sequences. The degradation on the background region is, however, relatively invisible to human perception in video telephony applications as illustrated in Fig. 2, which shows the per-frame PSNR comparison and the maximally improved frame on the face region and the maximally degraded frame on the whole frame.

Table I
Average PSNR comparison of the proposed DDWA, and TMN8 rate control schemes

| | PSNR of Frame (dB) | | PSNR of Face Region (dB) | |
|---|---|---|---|---|
| | 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 |

## 5. CONCLUDING REMARK

In this paper, we proposed a novel face detection and tracking scheme which can identify the face location using skin-color information in real-time. The skin-color pixels are segmented out by using a Gaussian probability model, and a low-complexity double integral projection method is subsequently used to segment out the face block from the pixels classified as the skin-color class. The proposed face detection and tracking algorithm can process up to 30 frames per second when implemented on a PC Pentium II 300 machine. We also proposed a dynamic distortion weighting adjustment method which skips the static non-face macroblocks to enhance the quality of face regions and reduce the computational cost. The experimental results show that the proposed algorithm can effectively enhance the visual quality of face regions at the cost of introducing some degradation on the non-face regions. The computational complexity of the proposed algorithm is pretty low thus making it well suited for real-time applications. Furthermore, the proposed algorithm is fully compatible to H.263 standard, thus can be integrated into the current commercial products.

## 6. 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 Communication*, Vol. 7, No. 4-6, pp. 231-248, Nov. 1995.

[2] S. Daley, K. Matthews, and J. Ribas-Corbera, "Face-based visually-optimized image sequence coding," *Proc. IEEE Int. Conf. Image Processing (ICIP), Chicago, IL*, pp. 443-447, Oct. 1998.

[3] 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, Jun. 1999.

[4] 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, Aug. 1997.

[5] 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.

[6] Image processing Lab, University of British Columbia, "H.263+ encoder/decoder," TMN codec, Feb. 1998.

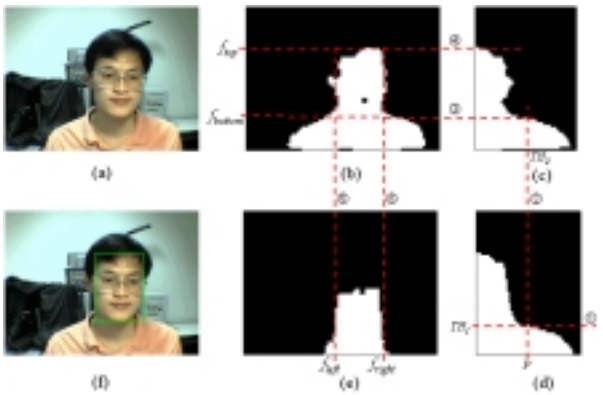
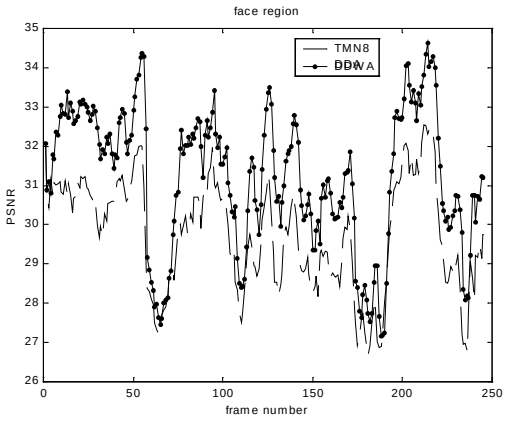
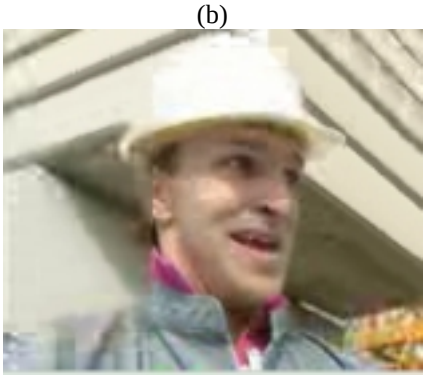


Fig. 1. The proposed double integral projection process for face detection (a) The original image, (b) The binary image generated from skin-color classification and processed by a median filter, (c) the horizontal integral projection image projected from (b), (d) the vertical integral projection image projected from (c), and (e) the vertical integral projection image projected from (b) between $f_{top}$ and $f_{bottom}$, (f) the detected face block accordingly.



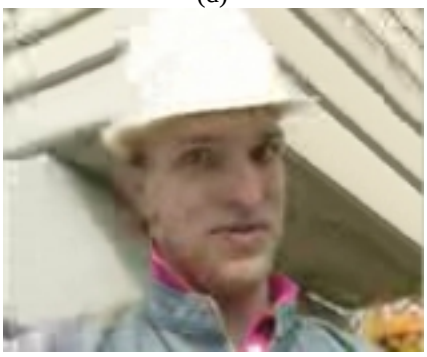

(a)

(b)



(c)

(d)



(e)

Fig. 2. (a) PSNR performance comparison of the proposed method with the TMN8 rate control on "Foreman" sequence at 64 Kb/s; (b) frame #224 coded using TMN-8 (32.32 dB); (c) frame #224 coded using the proposed method (34.72 dB0 (maximally improved frame on face region); (d) frame #201 coded using TMN8 (34.99 dB); (e) frame #201 coded using the proposed method (31.65 dB) (maximally degraded frame on whole frame).