A Multilevel Region-of-Interest Based Rate Control Scheme for Video Communication

QiLui Zhou*, Jiaying Liu, Zongming Guo
Institute of Computer Science and Technology, Peking University, Beijing, P.R. China

ABSTRACT

The ROI based video coding is widely applied in video communication. In this paper, we propose a multilevel ROI model, which includes the eye-mouth core region (CR), the face profile region (PR), the edge region (ER) and the background region (BR), to classify the subjective importance level of regions for the scene. Taking account of the proposed model, we first segment the current frame into four regions through skin color detection and feature location. Then, we improve the rate control algorithm in JVT-G012 proposal. We consider two factors, including subjective factor by our multi-level ROI model and objective factor by direct difference from reference frame, to model the complexity weight of each macroblock (MB).We allocate resources both at the frame layer and the basic unit layer, and adjust QP at MB layer. Finally, we restrict the QP of MB with three strategies to maintain the spatial and temporal smoothness. The experimental results illustrate that PSNR of ROI (CR plus PR) area using proposed method is in average over 0.5dB higher than JM8.6, while there are only slight changes in the PSNR of whole frame between two methods. Subjective quality based on our method also achieves much better performance.

Keywords: Region of Interest (ROI), Face Detection, H.264/AVC, Rate Control, Video Communication

1. INTRODUCTION

Recently, the ROI based video coding optimization is widely investigated with the development of video communication applications, including video telephone system, net meeting system, net talk show system, video interactive system, and remote education system, etc, which could use the achievements into industrial community, by promoting the image quality of video at a same bit rate as the former methods.

Several approaches have been proposed in this field, which can be divided into two categories based on the way of ROI segmentation. One is manual segmentation. Pietrowcew *et al.* [1] used level sets to denote local image complexity, and integrated with revised rate control method on ρ domain to reallocate resources. Luo *et al.* [2] focused on a hierarchical scheme of Flexible Macroblock Ordering (FMO) to encode the ROI and the non-ROI region in different slice groups and further divided the ROI region into several sub-slice groups. The other is automatic segmentation. Li *et al.* [3] used the position and motion information of a MB to define ROI, and employed a motion-based rate prediction model to realize bits allocation at the MB level. Liu *et al.* [4] utilized a face model as ROI. He uses pixel-level difference of the face areas to be his weight model. Simultaneously, a rate-distortion-complexity (RDC) cost function was proposed both at the encoder side and the decoder side.

In this paper, we propose a multi-level ROI model and correlative region segmentation method and rate control algorithm. Compared with the classical model taking face area as ROI, our model introduces a hierarchical subjective importance segmentation standard, and pays more attention to the feature areas. Taking account of this model, we utilize a skin-color and feature-based method to segment regions of frame, automatically. Then, an improved rate control algorithm is employed corresponding to reallocate resources and smooth the quality of different level regions.

The rest of the paper is organized as follows. A multilevel face-based ROI model is proposed in Section 2. An automatic image segmentation method based on the proposed model is described in Section 3. The improved rate control algorithm is discussed in Section 4. Experimental results are given in Section 5. Finally, concluding remarks are given in Section 6.

* towpence@163.com;

MIPPR 2009: Remote Sensing and GIS Data Processing and Other Applications,
edited by Henri Maître, Hong Sun, Bangjun Lei, Jufu Feng, Proc. of SPIE Vol. 7498,
74984W · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.833686

2. MULTILEVEL FACE-BASED ROI MODEL

We show the conventional definition of ROI for video communication in Fig. 1. Because there are a large number of frames mainly occupied with human faces in the scene, many methods divide the frame into two major parts: the face regions, which are drawn brighter in Fig. 1(b), and the background regions, which are drawn darker in Fig. 1(b).

(a) (b)

Fig. 1. Classical MB based ROI definition: (a) Original frame. (b) Classical MB based ROI definition. The brighter region is the collection of ROI MB.

However, we find that humans are more concerned about the facial features region, like eyes and mouth, and are less concerned about the profile of face relatively, like forehead, chin, and neck. Meanwhile, the border of face regions is more important than the background regions. So, we use a multi-layer-based region-of-interest model to distinguish the subjective importance of regions in frame. Fig. 2 shows the definition of the proposed model. Our model is based on MB, so as to be integrated into the video encoder more easily. Every square represents a MB. The MBs belonged to four regions are marked with different color, respectively. CR denotes the eye-mouth core region, PR denotes the profile of face region, ER denotes the edge of face region, and BR denotes the background region.

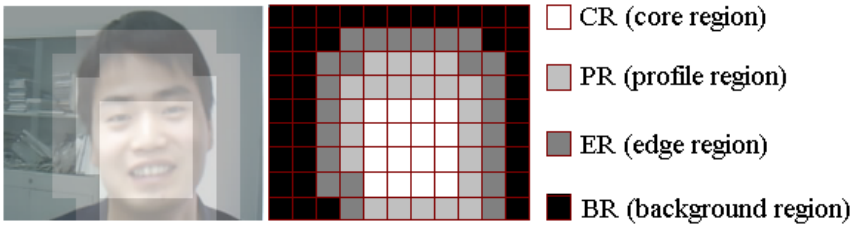


Fig. 2. Multilevel Face-Based ROI Model

Compared to the conventional ROI model, there are several advantages of the proposed model. Firstly, the model introduces the CR, PR, and ER, which is more accurate on the definition of subjective importance. Moreover, a hierarchical structure achieves better results in image quality smoothing. Finally, our model is MB based and is more convenient for video encoding.

3. SEGMENTATION

Our method of regions segmentation is composed of two major steps. First, we get the candidate face regions through skin color detection and a quick filtering. Then we locate the eyes and mouth using feature maps. The method is real-time, thus, it is used in video communication system.

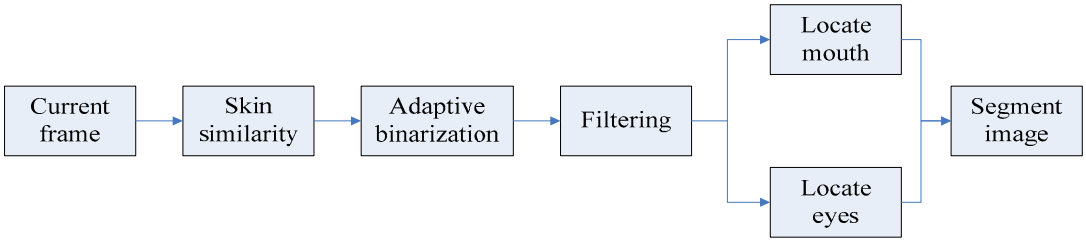


Fig. 3. Flow chart of the proposed algorithm

Fig. 3 shows the specific process of segmentation algorithm. In general, our method includes two parts. First we find the face region candidate by skin similarity computation, adaptive binarization, and filtering, then we locate the face features including eyes and mouth to segment the current frame into multilevel region of interest. We will introduce our method in detail in the following two sections.

3.1 Face region candidate

Many experiments show that the distribution of the *Cb* and *Cr* components of skin color is consistent with two-dimensional Gaussian distribution. So we adopt a single-peak Gaussian model (SGM) [5-7] defined below to compute the skin color similarity of pixel:

$$p(x/skin) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)\} \tag{1}$$

Where x is the vector (Cr, Cb), μ is the mean vector, $\sum$ is the covariance matrix, and both of μ and $\sum$ are estimated through statistical samples. The similarity result value is between [0, 1]. We use the following Equation (2) to find a threshold to segment the foreground and background.

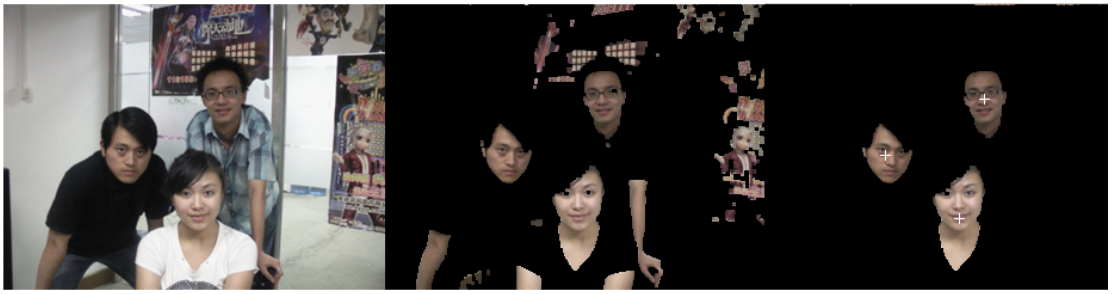
$$\arg\max_i \{w_b^i w_f^i (u_b^i - u_f^i)^2\} \tag{2}$$

Where i is the threshold, u_b^i and u_f^i stand for the average gray value of background and foreground regions with segmentation threshold i, w_b^i and w_f^i are the weight values, which are the proportion of background and foreground regions' size, respectively.

Then, we use feature filtering method to get candidate face regions. Two main face features are chosen to remove pseudo face regions. We first use the size feature by removing the foreground regions whose size is smaller than 3% of the whole picture or smaller than 25% of the largest connected foreground regions. Then, we use the shape feature, by removing the foreground regions whose long axis is 2.5 times longer than short axis in Equation (3).

$$\frac{\max(R_x, R_y)}{\min(R_x, R_y)} > 2.5 \tag{3}$$

where R_x and R_y are the amount of pixels which pass through the gravity center of the current foreground region in x-axis and y-axis, respectively. The candidate face regions after the filtering are shown in Fig. 4.



(a) (b) (c)

Fig. 4. Results of face region candidate detection. (a) Original frame. (b) Foreground regions after skin detection, where a lot of false regions are detected. (c) Filtering result, where three face region candidates are detected.

3.2 Features location

After getting the candidate face regions, we locate the eyes and mouth features to prepare for multi-level ROI segmentation. Li *et al.* [8] employed an eye map based on 314 eyes feature samples statistics. The eye map uses color feature to determine whether the pixel in candidate face regions is in eye region or not. Though Li's eye map could get the eye regions, it also gets many pseudo eye regions. So we improve the eye map generation method based on Equation (4), adding a factor to find the pixel whose SGM value is relatively low.,

$$Eyemap = \overline{sym}(Y - m_Y(k)) \quad and \quad sym(Cb - \max(100, m_{Cb}(k))) \\ and \quad \overline{sym}(Cr - \min(155, m_{Cr}(k))) \quad and \quad sym(S - \theta_s(k)) \tag{4}$$

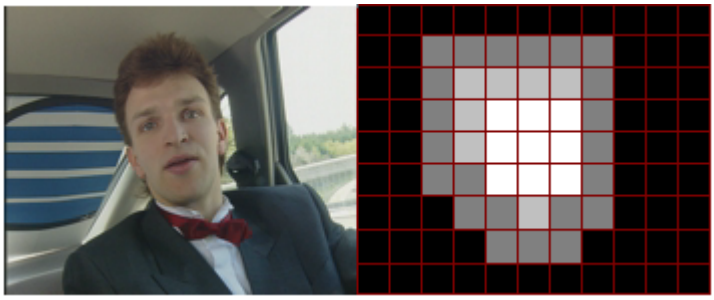
Where $sym(y) = \begin{cases} 1 & y \geq 0 \\ 0 & y < 0 \end{cases}$, $\overline{sym(y)} = \begin{cases} 1 & y < 0 \\ 0 & y \geq 0 \end{cases}$, $m_Y(k)$, $m_{Cb}(k)$ and $m_{Cr}(k)$ stand for Y, Cb and Cr mean value of the kth candidate face region respectively. S is the SGM value calculated before. $\theta_s(k)$ is the threshold determined by the S value of kth candidate face region. The eye detection result is shown in Fig. 5.

(a) (b) (c)

Fig. 5. Eye detection results. (a) Original frame of Carphone. (b) Result by Li's method. The eye regions are detected while a lot of false regions are detected as eyes regions. (c) Result by our method. The eyes regions are detected correctly, with few false regions.

With respect of other features of face, the human mouth is redder. Hsu [6] uses this to propose a mouth map based on *Cr* and *Cb* value. In this paper, we use Hsu's mouth map to get the mouth location.

Finally, we segment the frame into CR, PR, ER, and BR four regions after we get the face region, eyes region, and mouth region. Fig. 6 shows that the result of multi-level ROI segmentation.



(a) (b)

Fig. 6 Multi-level ROI segmentation result. (a) Original frame of Carphone. (b) Segmentation result, where blocks with different color represent CR, PR, ER, or BR, respectively, referring to Fig. 2.

4. MULTILEVEL ROI BASED RATE CONTROL

In this section, we will introduce our rate control algorithm based on our multilevel ROI model. First, we use two factors, including a subjective factor from multilevel ROI model and an objective factor from direct frame difference. In the

following Equation 5, we use to define coding complexity weight of the ith MB, which define the MB's complexity scale.

$$E_i = \alpha_i \times \sum_{(x,y)\in MB_i} \left(\left| f_{cur}(x,y) - f_{ref}(x,y) \right| \right),$$

$$\alpha_i = \begin{cases} 1.2 & MB_i \in CR_{cur} \\ 1.0 & MB_i \in PR_{cur} \\ 0.9 & MB_i \in ER_{cur} \\ 0.8 & MB_i \in BR_{cur} \end{cases} \tag{5}$$

Where f_{cur} and f_{ref} are the pixel value of current frame and reference frame. α_i is a threshold value which varies with CR, PR, ER, and BR. We use E_i to define coding complexity of the i[th] MB.

Then, we allocate bits for current frame and current basic unit through Equations (6) and (7). At the frame layer, we take into account the information of average bits for the rest frames and the coding complexity information of coded frames to decide how many bits should be allocated to the current frame.

$$B(i,j) = 0.5 \times \frac{B_{rf}(i,j)}{N_{rf}(i,j)} + 0.5 \times \left(\frac{\sum_{1<k<i} u(k,j)}{i-1} \times \sum_k E_k^i \right) \Bigg/ \frac{\sum_{1<p<i} \sum_k E_k^p}{i-1} \tag{6}$$

Where $B(i,j)$ is the bits amount allocated to the i[th] frame which is in the j[th] GOP, $B_{rf}(i,j)$ is the left bits amount of current GOP for current frame. $N_{rf}(i,j)$ is the number of frames not coded in current GOP. $u(k,j)$ is the actual bits for coding the k[th] frame which is in the j[th] GOP, E_k^i is the coding complexity of the i[th] frame which is in the k[th] MB, which is defined in Equation (5), and 0.5 is an empiric value.

At the basic unit layer, we allocate bits as following:

$$B(i,j) = B_{rb}(i,j) \times \sum_{k\in bu(i)} E_k^j \Big/ \sum_{k\geq i} \sum_{p\in bu(k)} E_p^j \tag{7}$$

Where $B(i,j)$ is the bits amount allocated to the i[th] basic unit which is in the j[th] frame, $B_{rb}(i,j)$ is the left bits amount of current frame for current basic unit, and $bu(i)$ is MBs' collection of the i[th] basic unit.

After we get initial QP0 for the basic unit by R-D model computation, we adjust QP of MB in current basic unit according to the ROI level of MB, defined in Equation (8):

$$QP_i = \begin{cases} QP_0 - 2 & MB_i \in CR \\ QP_0 - 1 & MB_i \in PR \\ QP_0 & MB_i \in ER \\ QP_0 + 1 & MB_i \in BR \end{cases} \tag{8}$$

Finally, the QP of MB is restricted with three parameters including the average QP of coded MBs which are in the same region level, the average QP of coded MBs (left, left-top, top, and right-top), and the average QP of previous frame. Each

parameter is used in order that the difference of QP is between 3. So the spatial and temporal smoothness can be well maintained.

5. EXPERIMENTAL RESULTS

We implemented experiments with the proposed algorithm with comparison to the rate control algorithm of reference codec JM8.6.The baseline profile is adopted, the coding format is "IPPP…". RDO is turned on, and the basic unit number is set to 11. We use three camera captured sequences, which are shown in Fig. 7, in addition with standard QCIF sequences.

(a) (b) (c)

Fig. 7. Camera sequence name. (a) Sequence M1. (b) Sequence M2. (c) Sequence M3.

The results of our method compared with JM8.6 are listed in Table 1.

Table 1 Result comparison with JM8.6

Sequence	Method	Rate(kbps)	fps	Avg. PSNR (dB)	ΔPSNR (dB)	ROI PSNR (dB)	ΔPSNR (dB)
Carphone	JM8.6	19.13	20	31.02	+0.20	30.10	+1.03
	Ours	19.14	20	31.22		31.13	
Foreman	JM8.6	30.41	20	29.43	-0.18	28.51	+0.39
	Ours	30.41	20	29.25		28.90	
Miss America	JM8.6	16.18	25	37.67	-0.27	33.57	+0.33
	Ours	16.28	25	37.40		33.90	
Claire	JM8.6	15.52	25	36.65	-0.15	31.74	+0.37
	Ours	15.44	25	36.50		32.11	
Suzie	JM8.6	25.35	25	32.20	+0.02	30.81	+0.10
	Ours	25.37	25	32.22		30.91	
M1	JM8.6	15.14	25	34.72	-0.26	34.11	+0.55
	Ours	15.16	25	34.46		34.66	
M2	JM8.6	15.13	25	31.27	-0.11	30.71	+0.58
	Ours	15.13	25	31.16		31.29	
M3	JM8.6	13.26	25	30.88	+0.11	29.92	+0.76
	Ours	13.32	25	30.99		30.68	

From this table 1, we draw the following conclusions. First of all, the bit rate difference between two methods is slight, and our rate control method is effective, which we could get from column rate. Besides, the average PSNR of the whole

frame between two methods is close from column Avg. PSNR. The most important conclusion is that PSNR of ROI (CR plus PR) using our method is 0.5dB greater than JM8.6 in average. Farther more, we find that our method gets better effect when the background is more complicated, like Carphone, M1, M2, and M3, also the effect of our method is weakened when the ROI is relatively big in the frame, like Suzie.

Fig. 8 shows the PSNR comparison results of Carphone sequence frame by frame. We can see that after 60th frame our method is more effective. That's because the coding complexity becomes greater.

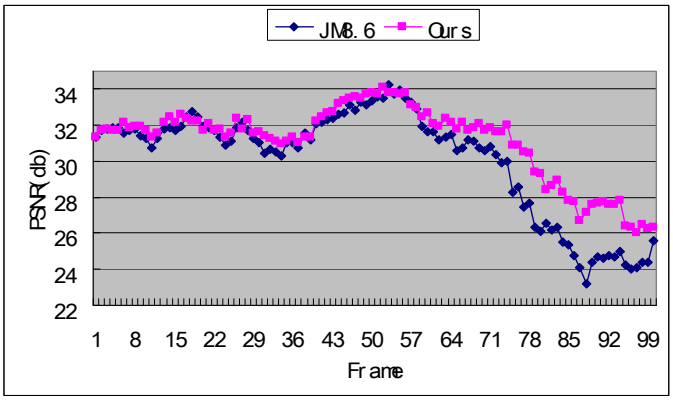


Fig. 9. Comparison frame by frame with JM8.6.

Fig. 10 shows the subjective result of consecutive frames (from 87th frame to 90th frame) in Carphone of two methods. The left column is JM8.6's result, while the right column is our result. We found that the face detail using our method is much clearer.

(a) JM8.6 (b) Ours (c) JM8.6 (d) Ours

Fig. 10. Subjective quality comparison with JM8.6. (a) Result of Carphone frame 87 and 88 by JM8.6. (b) Result of Carphone frame 87 and 88 by our method. (c) Result of Carphone frame 89 and 90 by JM8.6. (d) (c) Result of Carphone frame 89 and 90 by our method.

6. CONCLUSIONS

In this paper, we propose a multilevel region-of-interest model to weight the subjective importance of human. Based on this model, we propose an automatic segmentation method and a rate control algorithm. Our segmentation method uses some simply combined features to balance the accuracy and speed. We not only reallocate resource according to our multilevel ROI model, but also maintain the smoothness of visual quality. The experiment results show that our method achieves better PSNR in important regions than JM8.6, and get much better subjective quality. Our future work will focus on the integration of bits prediction model and our multi-level ROI model to predict the bits more precisely.

ACKNOWLEDGEMENTS

This work was supported by National Basic Research Program of China under contract No. 2009CB320907.

REFERENCES

[1] A. Pietrowcew, A.Buchowicz, W. Skarbek, "Bit-rate Control Algorithm Based on Local Image Complexity for Video Coding with ROI," Advanced Video and Signal Based Surveillance, 582-587 (2005).
[2] Rong Luo, Bin Chen, "A Hierarchical Scheme of Flexible Macroblock Ordering for ROI based 11.264/AVC Video Coding," ICACT, 1579-1582 (2008).
[3] Hang Li, Zhibing WangAn, et al, "Improved ROI-Based Rate Control Algorithm for H.264/AVC," Proc. ICSP, (2006).
[4] Yang Liu, Zheng Guo Li, Yeng Chai Soh, "Region-of-Interest Based Resource Allocation for Conversational Video Communication of H.264/AVC," IEEE Transactions on Circuits and Systems for Video Technology 1(18), 134-139 (2008).
[5] Cai J., Coshtasby A, "Detecting human faces in color images," Image and Vision Computing 18(1), 63-75 (1999).
[6] Hsu R. L., Abdel-Mottaleb M., Jain A., "Face detection in color images," IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 696-706 (2002).
[7] Pan Zhi-Geng, Zou Peng-Cheng, Liang Rong-Hua, "Human face detection using eigen-face and skin color," Journal of System Simulation 16(6), 1346-1349 (2004).
[8] Hongliang Li, King N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," Journal of Visual Communication and Image Representation 19(5), 320-333 (2008).
[9] Li Z., et al, "Adaptive Basic Unit Layer Rate Control for JVT," 7th Meeting: Pattaya Thailand, JVT-G012, (2003).