Ultra-Low Bit Rate Facial Coding Hybrid Model Based on Saliency Detection

Wei Zeng, Mingqiang Yang, and Zhenxing Cui
School of Information Science and Engineering, Shandong University, Jinan, China
Email: zw809791834@163.com, yangmq@sdu.edu.cn, 478715631@qq.com

***Abstract*—Aiming at getting high quality image sequences in real-time video chat applications at very low bit rate, we propose an improved model-based video coding system in this paper. Instead of detecting human faces with conventional ways, we detect salient regions in the frames using Boolean Map based Saliency (BMS) method to locate the faces. After facial feature extraction by AAM, we transmit the DPCM of the eyes and mouth directly instead of time-consuming facial expression estimation. The new approach we proposed is more suitable for real time and the simulation below indicates the proposed algorithm is effective and feasible.**

***Index Terms*—model-based coding, BMS, saliency detection, facial expression estimation**

I. Introduction

Recently, with the development of economy, social, technology, and the improvement of people's standard of living, the requirement for wireless communication, especially for video chat using cellphones or other communication tools, has increased. Although many researchers have done great efforts and propose a wide variety of methods in the respect of improving the transmission channel, the wireless internet is still at relatively low bit rate, which can negatively impact the user experience and limit the development of real-time video chat applications. Aiming to solve the problems caused by the limitation of low bit rate, we propose an improved ultra-low bit rate model for facial video coding/decoding on the basis of model-based coding (MBC) [1], [2] which has defined in MPEG-4.

The conventional coding methods such as predictive coding, transform coding, and vector quantization belong to information-theory based methods in which image signals are considered as random signals and compressed by exploiting their stochastic properties. These methods would suffer from severely blocky artifacts and mosquito effects at low bit rate. While model-based coding method, which represent image signals using structural image models, can still get high quality images even at very low bit rate. The major advantage of MBC is that it describes image content in a structural way and only needs to transmit the parameters of the model in the channel. Fig. 1 shows the general description of a model-based coding system.

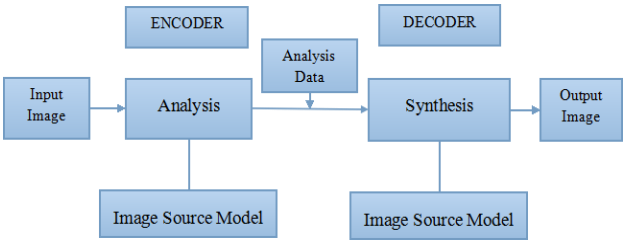


Figure 1. General description of a model-based coding system

But there are few applications using model-based coding for image coding/decoding of real time telecommunication up to now. One major reason is that the analysis process of model-based approach is too complex for real time. Detecting the human faces in the video frames using the existing methods costs too much time as a pre-processing of the model-based coding system. Estimating the facial expression using the facial action coding system (FACS) [3] also costs a lot of time and can only reconstruct approximate facial expression in the decoder.

Aiming at solving the problems above, we propose an improved model-based video coding system. We use a novel saliency detection method, Boolean Map based Saliency (BMS) approach [4], to detect the possible human face regions. And then using skin-color model for face verification. This approach is much faster than the existing human face detection approaches. In the transmitting procedure of the parameters which are needed to animate the model, instead of transmitting all the facial animation parameters defined in MPEG-4 [5], we only transmit the global motion parameters and the DPCM information of the regions of eyes and mouth. The changes of eyes and mouth can represent the facial expression in most cases and transmitting them directly can eliminate the time of facial expression estimation. What's more, transmitting the regions of eyes and mouth directly can get more realistic reconstructed images and more natural expression.

This paper is organized as follows. In Section 2, the selection of 3-D wire frame model is briefly introduced. Section 3 describes how to detect the face regions by BMS and face verification. Section 4 presents facial feature extraction and parameter estimation. Conclusions are described in Section 5.

Manuscript received February 13, 2015; revised June 15, 2015.

doi: 10.18178/joig.3.1.25-29

II. The Selection of 3-D Wire Frame Generic Face Models

Compared with some approaches which gradually build a model and dynamically modify it according to new video frames scanned, we use the 3-D wire frame generic face models in our proposed approach for the sake of simplicity and real-time. Another advantage of using generic face models is that these models are unrelated to any specific faces. The difference of all the human faces is the different value of the Facial Definition Parameters (FDP's) that transmitted in the initial frame. Once the value of FDP's has been received, the 3-D wire frame generic face model can specialize to the specific face in the video sequences. We choose Candide-3 [6] generic face model in our encoder and decoder after taking consideration of many aspects such as real-time, accuracy of details and sense of reality. Fig. 2 shows the 3-D wire frame generic face model we choose in our approach.

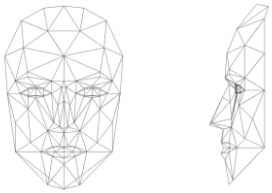


Figure 2. Candide-3 wire frame generic face model

III. Face Detection in the Initial Frame

Image analysis is the crucial part of our system. When there comes a frame of conversational video sequences, at first we need to analyze it and locate the human faces, Then we are able to extract the facial features and do further processing.

The relevant existing models usually detect and locate the faces in the image sequences using feature-based methods or appearance-based methods, such as skin-color feature detection [7], template matching, Adaboost algorithm [8], [9] and so on. But we think about it from a different perspective. Since we deal with the conversational video sequences, which means the sequences are under the assumption that all the frames of the sequences include at least one human face, instead of detecting a human face, we decide to detect visual saliency [10], [11] in the frames of the image sequences. We believe that the human face in the conversational video sequences ought to be salient. Actually, those frames which don't include human face can also be labeled by saliency detection method in our proposed model. Saliency detection approach can greatly reduce the computing time compared with the conventional face detection methods.

A. Locating Face Regions in the Frame Using BMS

Visual saliency detection has recently raised a great amount of research interest. There are so many different saliency detection methods. Many of them exploit the contrast and the rarity properties of local image patches to detect visual saliency, which is not suitable for our situation. We need to separate the human face from the background. We choose the Boolean Map based Saliency (BMS) approach [4] to detect the visual saliency in our system. BMS is based on the Gestalt principle of figure-ground segregation, which indicates that figures are more likely to be attended to than background elements, and the figure-ground assignment can occur without focal attention. That's why we believe that we can use BMS to detect the saliency of the conversational video sequences to find the potential human faces. Fig. 3 shows the pipeline of BMS.

The BMS model leverages the surroundedness of global topological cues for saliency detection. The essence of surroundedness is the enclosure topological relationship between the figure and the ground, which is well defined and invariant to various transformations. The BMS model firstly generates a set of Boolean maps by randomly thresholding the input image's color channels in CIE Lab color space. For each Boolean map, we compute an attention map based on a Gestalt principle for figure-ground segregation: surrounded regions are more likely to be perceived as figures. Assign 1 to the union of surrounded regions, and 0 to the rest of the Boolean map and then normalize the resultant map to gain the attention map. All the attention maps are linearly combined into a full-resolution mean attention map. Then we can process the mean attention map to the saliency map and segment the salient object, human faces in our case.

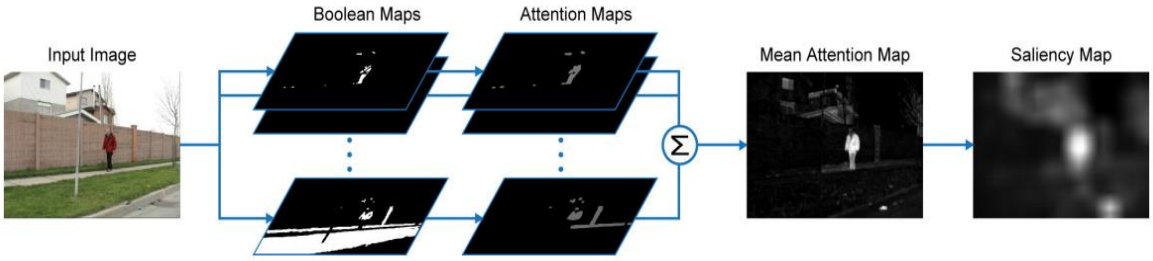


Figure 3. The pipeline of BMS

BMS is very simple to implement and efficient to run. The whole algorithm of BMS is summarized:

Step 1:

For each color channel map $\phi_k(I)\,(k=1,2,3)$ in Lab space of image I

For $\theta = 0:\delta:255$

$B = THRESH(\phi_k(I), \theta)$,

$\tilde{B} = \text{INVERT(B)}$

Add $OPENING(B, \omega_0)$ and $OPENING(\tilde{B}, \omega_0)$ to B

Step 2:

For each $B_k \in B$

$A_k = ZEROS(B_k, size())$

Set $A_k(i, j) = 1$ *if* B_k *belongs to a surrounded region*

$A_k = DILATION(A_k, \omega_{d1})$

$A_k = NORMALIZE(A)$

Step 3:

$$\bar{A} = \frac{1}{n}\sum_{k=1}^{n} A_k$$

Step 4:

$$S = POST\_PROCESS(\bar{A})$$

We use the BMS method to locate the human face regions in the frames of the conversational video sequences. Plenty of experiments have been done and the results indicate that our saliency detection approach is effective and feasible. Since we deal with the conversational video sequences which have relatively obvious salient regions, our recall rate can get to 95%. With the BMS method, we can greatly reduce the time cost by the conventional face detection process and match up our goal of real time. Fig. 4 shows some results of our approach. The first group is one frame of a video-chat sequence and we can locate the human face region quickly, even in some situations which have complex backgrounds, just as the second group shows, we can still detect the human face region using BMS successfully. But there may be some false regions in the original outputs of BMS, for example the window in the first group and the black chair in the second group, so we need further face verification to eliminate the false regions.

(a) The frame of one video-chat sequences and the output of BMS

(b) The frame with complex background and the output of BMS

Figure 4. Some results of the BMS method using in video sequences

B. Face Verification

After saliency detection, we should verify these candidate regions to ensure them to be the potential human faces. In our system, we use skin-color model [7] for verification.

Since the skin-color information is often affected by the color of the light source and the color deviation of equipment, illumination compensation should be done before skin-color verification. We use the "White Patch" approach to eliminate the possible color excursion, which is widely used because of its simplicity and effectivity.

The statistical skin-color model is generated by means of supervised training skin-color regions. Transform the training regions to YCbCr color space for better skin-color clustering and then use Gaussian model to represent the skin-color distribution and model skin regions. All these are done beforehand. We can use the skin-color model we obtained to verify the candidate regions to be human faces or not and eliminate the falsely detected regions. Fig. 5 below shows the results of using skin-color model for face verification. The first image of Fig. 5 is the segment of the BMS output, not only includes the face region but also has some false regions, after skin-color model verification, in the second image we eliminate the false regions and locate the human face region more accurately.

Figure 5. Face verification results using skin-color model

IV. Facial Feature Extraction and Parameter Extimation

Facial feature extraction is defined as the process of locating specific points or contours in a given facial image or aligning a face model to the image. In both cases, a set of two- or three-dimensional point coordinates is usually the output from the facial feature extraction. It's closely related to the parameter estimation process.

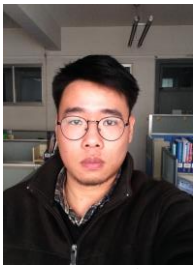
In a model-based video coding approach, three kinds of parameters need to be estimated: shape parameters, texture parameters and motion parameters. The shape parameters which describe the 3-D head shape are only transmitted in the initial frame. In MPEG-4 we use Facial Definition Parameter (FDP) to describe the shape parameters. The texture parameters describe the information that mapping to the face model. In MPEG-4 we usually use a waveform coder for the first frame and then only transmit difference images or no images at all for the following frames. The motion parameters include global motion and local motion parameters, which correspond to head motion and facial expression parameters, respectively. In MPEG-4 we use Facial Animation Parameter (FAP) to describe them.

A. Facial Feature Extraction

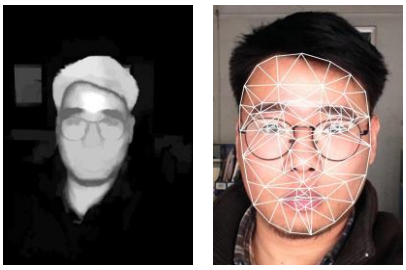
Having the candidate human faces, we need to locate the regions of eyes, noses and mouths and extract the facial features to specialize the 3-D wire frame generic face model to the faces in the input video sequences.

The Active Appearance Model (AAM) [12] algorithm has a profound mathematical background, excellent

characterization capabilities and it takes full advantage of a priori information about the object model. The AAM-based methods [13]-[15] have been widely used in facial feature extraction in recent researches. In our method we use the AAM method to locate the facial feature regions such as eyes, noses and mouths. The AAM method includes two parts: modeling and matching. The modeling process aims at mapping the information of human faces to the triangular mesh model generated from the training sets. The matching process adjusts the locations of the feature points on the model to match the faces by iterating the error function. After locating the facial feature regions by the AAM method, we can get the feature points we need to animate the 3-D wire frame generic face model. Fig. 6 shows the feature extraction result of AAM.



(a) The input frame



(b) Location of the candidate face (c) Output of AAM

Figure 6. The feature extraction of AAM

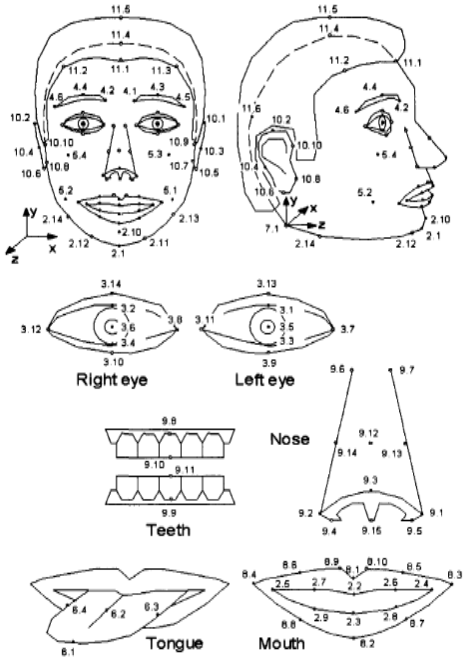


Figure 7. The feature points defined in MPEG-4

B. Parameters Estimation

After locating the facial feature points, we should extract the facial parameters for the 3-D wire frame generic face model. The model-based method defined in MPEG-4 video standard needs to extract the FDP's and the FAP's in encoder for the synthesis in decoder [16], [17]. While the facial expression estimation using Action Units of FACS is more or less time-consuming and unnatural after reconstruction. So we decide to improve the model-based method in MPEG-4 in this respect.

The MPEG-4 standard has defined 84 feature points to describe the human face model, including both FDP's and FAP's, which is showed in Fig. 7. Since we have located the feature points by AAM, we can directly extract the facial definition parameters as stipulated in MPEG-4 standard and specialize the Candide-3 generic face model in the initial frame.

The global motion estimation is to estimate the head motion parameters and reflect them to the 3-D wire frame face model. The global motion model is denoted as:

$$S' = RS + T \tag{1}$$

where R is the rotation matrix, T is the translation matrix, S is the position vector and S' is the corresponding position vector after motion. We use some of the feature points which wouldn't be affected by the changes of the facial expression, such as nasal tip, ears and so on, to calculate the rotation matrix and the translation matrix.

But as to the local motion parameters in FAP's, we don't use Action Units to describe the facial expression anymore. We transmit the DPCM of eyes and mouths instead since we believe that the changes in the regions of eyes and mouths can describe the facial expression changes in most cases. As the AAM has located the exact regions of eyes and mouths, it's easy for us to extract them and encode them using DPCM. What's more, since we transmit the regions directly, the reconstructed image can be more realistic and natural. Although transmitting the regions directly may increase the amount of data that need to transmit, it reduces the complexity of the facial expression estimation and become more suitable for real-time applications.

C. Bit Rate Estimation

The image sequences we use in our experiment are in the size of 640×480 and 30 frames per second.

The head motion is three-dimensional motion which has 6 parameters need to be transmitted [18], [19], 5 bit per parameter is enough, so the bit rate of the head motion parameters is:

$$M1 = 6 \times 5 \times 30 = 900 \ \ bit$$

The bit rate of the DPCM of eyes and mouth depends on every different frame. In our experiment the average update of the facial expression need about 24500 bit per second.

So the average bit rate for a frame is 25.4 kbps. It indicates that our approach can conquer the ultra-low bit rate situation.

V. Conclusion

In this paper we propose an improved model-based coding system. We use BMS to detect the salient regions in conversational video sequences instead of the conventional face-detecting approaches to locate the human face regions in order to reduce the computing time of the pre-process. We transmit the DPCM of eyes and mouth instead of the time-consuming facial expression estimation. Our proposed approach can transmit high quality facial image sequences in real time through ultra-low bandwidth on wireless channel.

Acknowledgements

The research work was supported by Natural Science Foundation of Shandong Province under Grant No. ZR2014FM030 and No. ZR2013FM032.

References

[1] R. Forchheimer and T. Kronander, "Image coding-from waveforms to animation," *IEEE Trans. Acoustic Speech Signal Processing*, vol. 37, no. 12, pp. 2008-2023, 1989.
[2] K. Aizawa, H. Harashima, and T. Saito, "Model-Based analysis synthesis image coding (MBASIC) system for a person's face," *Signal Processing: Image Communication*, vol. 1, no. 2, pp. 139-152, 1989.
[3] P. Ekman and W. V. Friesen, *Facial Action Coding System*, Palo Alto: College Arc Consulting Psychologists Press, 1977.
[4] J. M. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2013.
[5] S. Bauer, J. Kneip, T. Mlasko, *et al.*, "The MPEG-4 multimedia coding standard: Algorithms, architectures and applications," *Journal of VLSI Signal Processing*, vol. 23, pp. 7-26, 1999.
[6] J. Ahlberg, "CANDIDE3-An updated parametrized face," Technical Report No. LiTH-ISY-R-2326, Department of Electrical Engineering, Sweden, 2001.
[7] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, pp. 1106-1122, 2007.
[8] P. Viola, "Rapid object detection using a boosted cascade of simple features," in *Proc. International Conference on Computer Vision and Pattern Recognition*, 2001, vol. 4, pp. 624-627.
[9] C. Zhang and Z. Y. Zhang, "A survey of recent advances in face detection," Microsoft Technical Report, MSR-TR-2010-66, Jun. 2010.
[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency based visual attention for rapid scene analysis," *PAMI*, vol. 20, no. 11, pp. 1254-1259, 1998.
[11] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-Tuned salient region detection," in *Proc. CVPR*, 2009.
[12] J. Ahlberg and R. Forehheirner, "Face tracking for mode-based coding and face animation," *International Journal on Imaging Systems and Technology*, vol. 12, no. 1, pp. 8-22, 2003.
[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. European Conf. Computer Vision*, 1998, vol. 2, pp. 484-498.
[14] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135-164, 2004.
[15] M. B. Stegmann, "Active appearance models: Theory, extensions and cases," IMM (Informatics and Mathematical Modeling), Technical University of Denmark, 2000.
[16] C. G. Yang, W. W. Gong, and L. Yu, "Automatic facial animation parameters extraction in MPEG-4 visual communication," in *Proc. SPIE-The International Society for Optical Engineering*, California, USA, 2002, pp. 396-405.
[17] L. Yu, J. Y. Mang, and W. W. Gong, "Parameter analysis and synthesis for MPEG-4 facial animation," in *Proc. Workshop and Exhibition on MPEG-4*, 2001, pp. 49-52.
[18] J Ahlberg, "Model-Based coding: Extraction, coding, and evaluation of face model parameters," Ph.D. thesis, Linköping University, Sweden, 2002.
[19] M. Kampmann, "Automatic 3-D face mode adaption for model-based coding of videophone sequences," *TCSVT*, vol. 12, no. 3, pp. 172-182, 2002

Wei Zeng received the B.E. degree in School of Information Science and Engineering, Shandong University, Jinan, China, in 2013. He is currently a master student in image processing laboratory at Shandong University. He will receive his master's degree in June, 2016. His research interests include image processing, pattern recognition and computer vision.

Mingqiang Yang received his M.Sc. in Signal Processing from Shandong University in 2000. He joined the National Institute for Applied Sciences of Rennes (INSA) in France for his PhD studies in Image Processing and received this degree in 2008. He is associated professor in Shandong University. His research interests include image analysis, pattern recognition, image fusion and remote sensing.

Zhenxing Cui received the B.E. degree in College of Physics and Electronics, Shandong Normal University, Jinan, China, in 2013. He is currently a master student in image processing laboratory at Shandong University. He will receive his master's degree in June, 2016. His research interests include image processing and feature extraction.