2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

An Adaptive Face Model for Real-Time Eyes and Mouth Tracking

Shahrel A. Suandi [1,2 +], Shuichi Enokida [2] and Toshiaki Ejima [2]

[1] School of Electrical & Electronic Engineering, Engineering Campus, Universiti Sains Malaysia,
14300 NibongTebal, Pulau Pinang, Malaysia.

[2] Intelligent Media Laboratory, Department of Artificial Intelligence, Kyushu Institute of Technology,
Iizuka City, Fukuoka Pref., Japan.

Abstract. We introduce an adaptive face model to be used with extended template matching (ETM) technique [1] to track eyes and mouth in real-time. Although ETM is shown to be robust for these facial components detection, its' performance reduces when the face is not facing forward. This is due to failure of determining correct face center during the tracking, especially from non-frontal pose. Our proposed adaptive face model is meant to resolve this problem. It is simple and reliable to track these facial components within approximately 70° toward left and right. By using face geometrical relationships between both eyes, we create a face model adaptively by dividing the detected face region into right and left regions based on discriminant analysis method. This is done continuously in each frame to create the adaptive face model. It works as the base model during facial components searching using ETM. Results reveal that the proposed method performs better than [1] and tracks at averaged rate of 30 fps.

Keywords: template matching, eyes and mouth tracking, model-based tracking

1. Introduction

In the framework of tracking eyes and mouth in real-time, there are a lot of techniques have been proposed. These include intrusive and non-intrusive techniques depending on how the systems requirements are. As for the methods, they can be broadly categorized into knowledge-based, feature invariant, e.g. texture, skin colour, shapes, size, etc., appearance-based methods, e.g. eigenface, SVM, Bayes Classifier, etc., and template matching [2]. To certain extent, most of these methods show satisfactory results but however, we are not aware of methods that can track the facial components (including the area) in real-time as long as they are still in the view of the camera. This factor is important for applications like human-robot interaction, face gesture recognition, and monitoring driver alertness because it is necessary to let the subject moves freely, which obviously will sometime make the subject distracts their view from the camera.

In previous works that have been published by Tian et al. [3], they address out problems to track eyes employing deformable template matching, and feature point tracking. The former is claimed to have disadvantages of time consuming and difficult to be implemented in image sequences, while the latter is claimed to produce errorness when the eyes blink. To overcome the error when blinking, they proposed dual-state model which consists of open and close models so that the tracker will know what to do if a model is detected. In contrast to this approach, Funayama et al. [4] successfully extracted facial components, i.e. eyes and mouth, using a method given by a name of active nets [5]. This approach utilizes active contour model (snake) from deformable template technique and face geometrical constraints about the placement of the corresponding facial components. The method is robust to rotated images. On the other hand, similarly to deformable template technique, a technique that utilizes predefined face templates [2] (in this paper, we refer

[+] Corresponding author. Tel.: + 60 4 5995814; fax: +60 4 5941923.
E-mail address: shahrel@eng.usm.my

to this technique as template matching(TM) for simplicity), is also broadly used. It is a typical method for facial component extraction [4], [6] and has about 30 years of historical records for computer recognition of human faces. Despite of its simplicity, it suffers from lighting condition, background noises and weak towards rotation. Because of these reasons, it is not applicable directly in tracking framework [7]. Works employing TM as the principal technique will either use it at initialization stage [8], or small and limited window size at the interested area after being localized during initialization [7], [9]. Gargesha et al. [10] propose a hybrid method using combination of a few computer vision approaches along with TM. We have also managed to implement TM in real-time [1] using template matching technique but in extended manner, or known as "Extended Template Matching (ETM)". However, the performance decreases when users are not at straight frontal forward pose.

In this paper, we describe the solution to the problem faced in [1] in order to increase the tracking performance. Our contribution in this work is to design a simple, reliable and non-intrusive eyes and mouth tracking system that is able to track these facial components as long as they exist in the view of the camera. The tracker works by utilizing the center position of both pupils from previous frame as the separator to divide the detected face region into two regions; right and left. This becomes our face model during tracking and the reference where to search the corresponding facial components. By creating a face model adaptively like this, we can estimate the center of face while at the same time estimate where the positions of both eyes and mouth are. Locations of these facial components are important for other systems development [11], [12]. In TM technique framework, having a reference like the face model will help and support the system in the perspectives of speed and accuracy. For eyes, mouth and pupils detection, we use similar method as being proposed in [1]. Through experiments, we have observed that our tracker is capable to handle tracking until approximately $70°$ to the right and left with high detection accuracy for eyes, mouth and pupils. The remainder of this paper is organized as follows. We briefly describe the technique of extended template matching [1] in Section 2 and in the subsequent section, we describe generally how the face is being located in the video sequence. In Section 4, we present in details our proposed method. Experiments and the results are shown in Section 5 and finally, we put our conclusion in Section 6.

2. Extended Template Matching

Extended template matching (ETM) is a method utilizing template matching(TM) as the principal technique. But, due to the limitations in TM, we employ more than one higher correlation value for each corresponding facial component (instead of employing the maximum correlation value) to determine the possible area. Selections for the best candidates for both eyes and mouth are made based on two selective functions shown in Eq. (1) and Eq. (2), respectively. Both equations seek for the candidates that minimize the total of energy among them based on energy minimization criterion (EMC). Selection done in this manner will reduce dependency on correlation values, which therefore, suitable for tracking a non-static object that has some fixed geometrical constraints, for instance, eyes and mouth. Both Eq. (1) and Eq. (2) are defined as:

$$\mathrm{S}_{\mathrm{eyes}}(i,j) = \lambda_1 f_y(i,j) + \lambda_2 f_x(i,j) + \lambda_3 f_{LC}(i) + \lambda_4 f_{RC}(j), \tag{1}$$

$$\mathrm{S}_{\mathrm{mouth}}(i) = \lambda_1 \Delta\mathcal{M}_i + \lambda_2 f_m(i) + \lambda_3 f_{MC}(i), \tag{2}$$

where, all $\lambda_k\ (k=1,2,3,4)$ in both equations are the control parameters, f_y is a vertical distance between both eyes candidates, f_x is a horizontal distance between both eyes candidates, f_{LC} is a normalized left eye correlation value, f_{RC} is a normalized right eye correlation value, ΔM_i is a vertical distance between selected eyes and mouth candidates, f_m is a horizontal distance between center of face and center of mouth candidates, f_{MC} is a normalized mouth correlation value, (i,j) in S_{eyes} denotes an index for corresponding candidate number in left and right eyes candidates respectively, and (i) in S_{mouth} denotes an index for corresponding candidate number in mouth candidates. Each of f_{LC}, f_{RC} and f_{MC} is normalized by deducting the correlation value from 1.0.

3. Face Localization

In this section, we present our method to detect and locate a face from a video sequence. Methods to extract human face from an image have been published in [13], [14], [15], [16], [17] using color information, and [18], [19] using statistical learning approach. An overall survey for existing methods is reported by Yang et al. [2]. Generally, approaches using human skin-color distribution are preferred due to its' computational simplicity reason and robustness towards complex scene. In such approach there are two techniques exist; (1) determine the distance to a distribution – human-skin color model is prepared beforehand and the distance from a pixel in local image to the distribution model using Euclidean, Mahalanobis [8], [4] distance, etc. is computed to determine which class the unknown pixel attributes in, and (2)define a value range as the parameters for each domain in a color space through experiment [16].

In our approach, we locate skin-like regions by performing color segmentation and verify the parameters empirically. For segmentation purpose, we consider the *rg color space* and refer to work done by Terrillon et al. [15]. They show that normalized *rg* color space is independent from camera types and robust to extract human skin-color. Alternatively, similar color spaces (*e.g.* HSV, TSL, YIQ) can also be used. For skin-color segmentation, it is sufficient to consider r and g as discriminating color information. These domains describe the human skin-color and can be defined or estimated a priori to be used as the reference for any human skin color. In our work, we have defined the parameters for each domain as follows: $0.37 \leq r \leq 0.47$ and $0.28 \leq g \leq 0.35$. In Fig. 1(b), we show an example of our segmentation method.

To verify and locate where the face is, we make an assumption that the biggest region exists in the image is the face. At first, we make regions by connecting pixels that are 8-neighbor connected and subsequently, delete small regions by applying background dilation algorithm, as shown in Fig. 1(c), (d). Then, we compute the vertical and horizontal projections using this result image to estimate the location of the face vertically and horizontally (Fig. 1(e)).

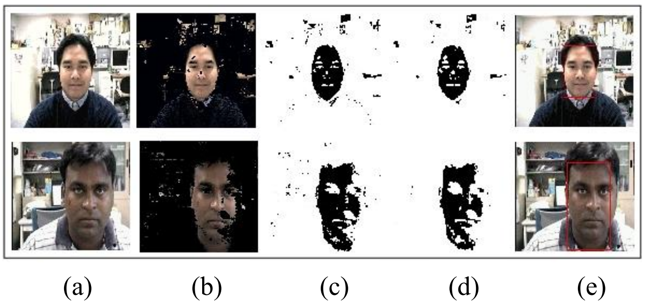

Fig.1: Detection of face region: (a)input image, (b)skin-color pixels, (c)after being binarized and region segmentation, (d)after background dilation, and (e)the face region in rectangle

4. Adaptive Face Model

Face model is important in our approach because ETM results are dependent on this model for the geometric relationship. As mentioned in Sect. 2, besides the correlation values, information based on the global position of corresponding facial components are also being considered in ETM. In other words, candidates that minimize the total energy are actually placed with a certain geometric constraint on the face and the most important thing is they are all attracted to the center of the face. Thus, information such as "at the center below of both eyes, there exists a mouth" is appropriate and can be used as the prior knowledge where to search the facial components. As a matter of fact, in the real world, this is always applicable whenever head rotates to the right or left. Therefore, it is necessary to know where the "center" is in each frame. To realize this, we determine a value to separate the face region into two regions horizontally. This value is given by a name; *face center line*. Having this value changed accordingly when head rotates to left or right, we achieve an adaptive face model which is shown to be robust to head horizontal rotation.

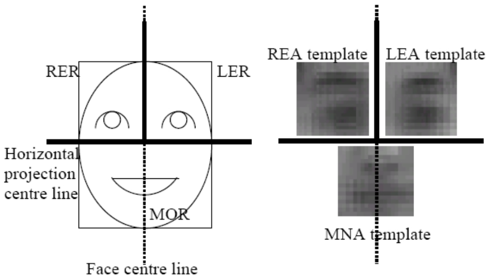

Fig. 2: Initial face model showing corresponding regions in our work and the situation when total energy is the minimum

4.1 Face Model

The face model in our work is defined as a model consists of a few regions that are known as right eye region(RER), left eye region(LER) and mouth region(MOR)(shown in Fig. 2). Interested areas, which are areas of right eye (REA), left eye(LEA) and mouth nose(MNA) will be searched within these RER, LER and MOR, respectively. Regions in face model are fixed by verifying two values; i.e. horizontal projection center line and face center line. The former divides the face region vertically, while, the latter divides the face region horizontally. However, only face center line is used during tracking. Due to this difference, we associate two types of face model in the tracker. The definitions of each of them are given below:

- **Initial face model** – this model is used only during initialization, which occurs either when the tracker begins tracking or when it recovers itself due to error detection. It consists of RER, LER and MOR.
- **Adaptive face model** – this model is renewed in each frame and is applied only during tracking. It consists only RER and LER. When the position of both pupils are known in frame t, the area between both pupils at this time t will be utilized to define the face center line for the subsequent frame at time t +1. Searching for REA, LEA and MNA candidates will be performed within a small specified area after being localized in the previous frame (at time t). Therefore, MOR is not required during tracking.

4.2 Creating Adaptive Face Model

In creating both face models, at first, we transform the face region into an image that consists horizontal facial components by using horizontal haar wavelet transformation. Other edge detection methods can be used as well but haar wavelet is preferred in our work due to its' advantage for multi resolution analysis. Then, we compute the vertical and horizontal projections within the face region. Using discriminant analysis method, we calculate the thresholds for both projections. The vertical projection's threshold is utilized as the face center line value and the horizontal projection's threshold is utilized as the horizontal projection center line value. From these two values, an initial face model is created. ETM is applied after this to select correct eyes and mouth areas. Pupils are searched within the selected REA and LEA.

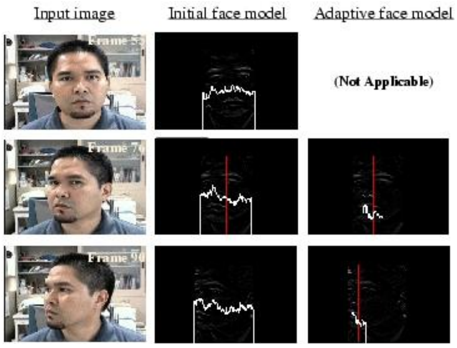

Fig. 3: Vertical projection using different face model. Adaptive face model shown to be more reliable than the initial face model in handling face pose.

Utilizing only the initial face model is insufficient in a facial features tracking framework. The face model is built by analyzing vertical and horizontal projections. Both projections rely solely on horizontal facial components that have been extracted during the transformation. As a result, when the face rotates to the right or left, some horizontal components such as edges between hair and forehead, ears and back necks, influence the projections especially the vertical projection. This has affected the face center line as well, which finally results in false detections for all facial components. Fig. 3 illustrates these problems. The red line shows the face center line. Therefore, an adaptive face model is desired to overcome this problem.

Since considering the face region is not relevant and the center of the face is where the facial components rely on, we utilize region between both detected pupils to calculate the face center line. We illustrate this in Fig. 4. Even though both results show possibilities to be feasible for creating the adaptive face model, we choose the upper method because it is simpler and assumed to content lesser noise if compared to the lower method.

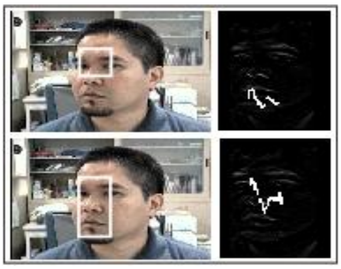

Fig. 4: Vertical projection results from two different areas in rectangle; images on left show the area while images on right show the actual projection results

4.3 Tracking using ETM

Tracking related facial components are done after initialization. At initialization stage, the tracker will find REA, LEA and MNA using ETM technique and following this, pupils area are searched within the selected REA and LEA using normal template matching. Actual pupils are defined as the darkest pixel in the pupils' areas. When these facial components are found, we record their positions. During tracking we search for these facial components by searching at small area around the previously detected positions of these components. As for selection for the correct candidates for REA, LEA and MNA, we keep using ETM technique, while for the pupils; we search for the darkest pixel at previously detected positions.

5. Experimental Results

For evaluation purpose, we have prepared two video sequences consist of about 600 images each. The image size is 200 × 300. Each video sequence consists a single face in our laboratory environment with uniform lighting. The persons were asked to rotate their faces to right and left until one of the eyes is mostly disappear from the camera's view. We set this situation as approximately 70°. They were also asked to move their faces (without rotation, i.e. while looking forward) to right and left. The experiment is done on a FreeBSD4.7 OS machine with a Celeron 2.2 Ghz CPU and 512MByte of memory. The processing speed is 30 frames per second (fps).We show some of the tracking results for both persons in Fig. 5. For the first person, we have observed a 98% tracking accuracy for all facial components, but for the second person, about 90% tracking accuracy has been observed. More false detection results are observed on the second database set due to the person rotates his face more than required, i.e. approximately $80^\circ \sim 90^\circ$ (see frame 460, 470, 540 and 548 in Fig. 5). The tracker will keep on tracking until it encounters error based on some rules that have been set beforehand. This explains why in the frames 460, 470, 540 and 548 the tracker still displays false results. When an error is detected, it will recover itself automatically starting from the initialization stage. However, while recovering itself, initial face model will be used as the base model to search for the related facial components. This would result an error as well, if the face remains at the same pose as it was or at any pose that is greater than approximately 40° to the right or left. This remains as our future work where during initialization, it is essential to recognize the face pose to resolve this particular problem.

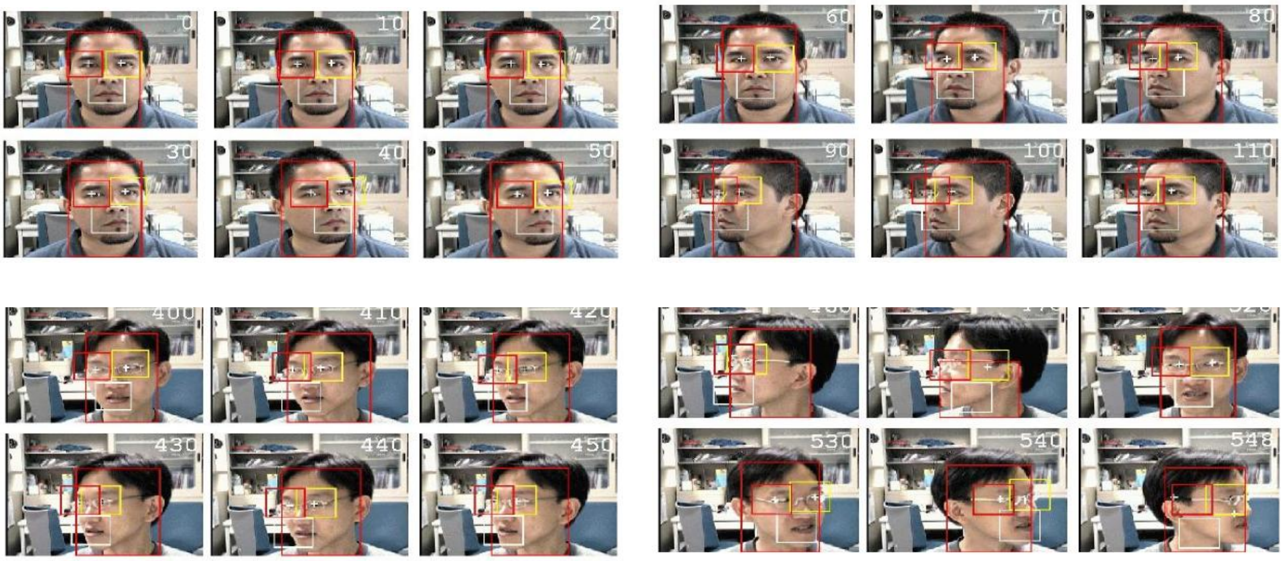

Fig. 5: Tracking results by our method for subjects with varying face pose.
On the top right is shown the frame number

6. Conclusion and Discussion

We have presented a simple and reliable method to track eyes and mouth from non-frontal face using an adaptive face model as the base model for extended template matching implementation. The method is a non-intrusive method which is very convenient for a man-machine interactive communication system and can be operate data video processing rates. Utilizing two face models, which are initial face model as the model to locate eyes and mouth, and adaptive face model as the model during tracking, the system is able to locate and track a user's eyes, mouth and pupils as soon as the user appears in the view of the camera. This is done fully automatic without any initialization or calibration, e.g. light adjustments. The system has achieved accuracy greater than 90% for two data sets in the experiment. However, for the cases where face rotates more than 70° to the right or left, the system fails. This is actually caused by the failure of the face model. To overcome this problem, we need some additional information for the system to be able to recognize the face situation while it tracks. We will address this problem in our further research.

7. Acknowledgement

This work is partially supported by Universiti Sains Malaysia ShortTerm Grant (304/PELECT/6035270).

8. References

[1] S. A. Suandi, S. Enokida, and T. Ejima, "An extended template matching technique for tracking eyes and mouth in real-time," in *Proceeding 3rd IASTED Int'l Conf. Visualization, Imaging and Image Processing*, 2003.

[2] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images:Asurvey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.

[3] Y. L. Tian, T. Kanade, and J. F. Cohn, "Dual-state parametric eye tracking," in *4th IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 2, March 2000, pp. 110–115.

[4] R. Funayama, N. Yokoya, H. Iwasa, and H. Takemura, “Facial component extraction by cooperative active nets with global constraints,” in *13th IEEE International Conference on Pattern Recognition (ICPR)*, vol. 2, 1996, pp. 300–305.

[5] K. Sakaue and K.Yamamoto, “Active net model and its application to region extraction,” *The Journal of the Institute of Television Engineers of Japan*, vol. 45, no. 10, pp. 1155–1163, 1991, in Japanese.

[6] R. Brunelli and T. Poggio, “Face recognition: Features versus templates,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 992–1006, October 1993.

[7] J. Heinzmann and A. Zelinsky, “Robust real-time face tracking and gesture recognition,” in *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI’97*, vol. 2, 1997, pp. 1525–1530.

[8] R. S. Feris, T. E. de Campos, and R. M. C. Junior, “Detection and tracking of facial features in video sequences,” *Lecture Notes in Artificial Intelligence*, vol. 1793, pp. 197–206, April 2000.

[9] Y. Matsumoto and A. Zelinsky, “Real-time face tracking system for human-robot interaction,” in *9th IEEE International Conference on Systems, Man and Cybernetics (SMC’99)*, vol. II, Oct, 12-15 1999, pp. 830–835.

[10] M. Gasgesha and S. Panchanathan, “A hybrid technique for facial feature point detection,” in *Fifth IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI”02)*, 2002.

[11] S. A. Suandi, S. Enokida, and T. Ejima, “Horizontal human face pose determination using pupils and skin region positions,” *Lecture Notes in Computer Science*, pp. 413–426, December 2007.

[12] S. A. Suandi, S. Enokida, and T. Ejima, “Face pose estimation from video sequence using dynamic bayesian network,” in *IEEE Workshop on Motion and Video Computing*, January 2008.

[13] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, “Face detection in color images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, May 2002.

[14] M.-H.Yang and N. Ahuja, “Detecting humanfaces in color images,” in *Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP 98)*, vol. 1, October, 1998, pp. 127–130.

[15] J.-C.Terrillon, A. Pilpret,Y. Niwa, and K.Yamamoto, “Properties of human skin color observed for a large set of chrominance spaces and for different camera systems,” in *8th Symposium on SensingVia Image Information*, 2002, pp. 457–462.

[16] K. Sobottka and I. Pitas, “Extraction of facial regions and features using color and shape information,” in *International Conference on Pattern Recognition (ICPR)*, vol. III, 25–29 August 1996, pp. C421– C425.

[17] J.Yang,W. Lu, and A.Waibel, “Skin-color modeling and adaptation,” Carnegie Mellon University,Tech. Rep. CMU-CS-97-146, May 1997.

[18] E. Osuna, R. Freud, and F. Girosi, “Training support vector machines: an application to face detection,” in *Proceedings of Computer Vision and Pattern Recognition*, 17-19 June 1997.

[19] A. J. Colmenarez, “Facial analysis from continuous video with application to human-computer interface,” Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1999.