EMoTracker: Eyes and Mouth Tracker Based on Energy Minimization Criterion

Shahrel A Suandi, Shuichi Enokida and Toshiaki Ejima
Intelligence Media Laboratory, Kyushu Institute of Technology,
Iizuka City, 820-8502 Fukuoka Pref., Japan.
{shahrel,enokida,toshi}@mickey.ai.kyutech.ac.jp

Abstract

*We introduce a novel approach for online facial components tracking based on energy minimization criterion. The tracker, known as **EMoTracker**, employs template matching as the principal technique. As feature appearance changes during tracking, template matching suffers in providing good detection results. Therefore, instead of utilizing only the similarity (correlation values) independently, we add global constraints of facial components placement on face as additional parameters when searching corresponding components. In order to define the correct areas, we first list out n areas (which are the candidates) for eyes and mouth employing template matching technique. These candidates are arranged in high to low correlation order. Selections of correct candidates among these candidates are made based on energy minimization criterion. Additionally, an automatic feature selector and an adaptive face model have been incorporated with EMoTracker to handle tracking from multiple type of faces and non-frontal faces, respectively. Our proposed method also requires no manual initialization and parameter tuning.*

1. Introduction

In the framework of tracking eyes and mouth in real-time, there are lots of techniques have been proposed. These include intrusive and non-intrusive techniques depending on how the system requirements are. As for the methods, they can be broadly categorized into knowledge-based, feature invariant (*e.g.* texture, skin color, shapes, size, etc.), appearance-based methods (*e.g.* eigenface, SVM, Bayes Classifier, etc.), and template matching (TM) [17]. We employ TM as the principal technique in our work. Despite of its simplicity, for implementation, it suffers from lighting condition, background noises and weak towards rotation. Because of these reasons, it is not applicable directly in tracking framework [4]. Works employing TM as the principal technique (in tracking framework) will either use it only during initialization [2] or small and limited subwindow size at the interested area after being localized during initialization [4][6].

In contrast with these ancestors methods, we apply TM in each frame. The facial components that we are interested in are eyes, mouth and pupils. Being able to track these facial components at high detection accuracy and speed, we are looking forward to utilize this tracker for some other applications, for example, head gestures recognition, facial expression for security, human-robot interactive communication, etc. The principal idea in this work is to combine image correlation results with face geometrical constraints of facial components placement on face. Assumption that "when eyes and mouth are placed correctly according to some face geometrical constraints, the total energy is minimized" is used to define correct components. These components are said to have *energy minimization criterion (EMC)* when they comply with this assumption. In order to define how the energy is minimized mathematically, we introduce two functions that take into consideration the correlation values and face geometrical constraints. We also prepare a *face model* – a model built based on face center (vertical and horizontal) lines. This model is utilized as the TM platform. On top of that, as face is not a static object, we make the model changes adaptively during tracking. The contribution of this work is to provide a method to track not only eyes and mouth but the surrounding area as well, as long as they exist within the view of a camera. As the result, the tracker can be widely used by various applications and therefore, allows one to use information from these areas for other applications.

The approach presented in this paper has similar concept with the work proposed by Funayama et al.[3] and Smith et al.[11]. The former has successfully extracted facial components (regardless face orientation) using a method known as *cooperative nets*, which comes originally from a method proposed by Sakaue et al.(*active nets*)[9]. We adopt the idea of mouth center is always perpendicular to the line con-

necting the center of both eyes to select mouth from their work. This creates robustness towards face clockwise or anti-clockwise rotation (roll action). The latter reports a tracking method to track pupils and mouth corners to monitor driver alertness. We use similar method in our work when tracking pupils but however, for eyes detection, TM is employed. To realize this approach as an online application, we reduce the area for TM by employing color information to detect human skin-color pixels and create face area from the detected pixels. The largest skin-color region is considered as the face area. Horizontal size (width) of this face area is utilized to estimate size of corresponding facial components. By doing this, the tracker is capable of adjusting the patch size automatically during TM when the tracked subject moves nearer or further from the camera. After detecting and localizing each facial components, tracking is performed by searching at previously detected position. In order to track from non-frontal pose, an adaptive face model is proposed. Results reveal that our method shows improvements in terms of time and accuracy compared to our previous reported work [13].

The remainder of this paper is organized as follows. We briefly describe how the face is being detected and located from the video sequence in Section 2. In Section 3, we describe the adaptive face model and in the subsequent section, we present our main method using energy minimization criterion(EMC). Experiments and the results are shown in Section 5. Finally, we put our conclusion in Section 6.

2. Face Detection and Localization

In this section, we present our method to detect and locate a face from a video sequence. This is the most important and critical task in our work because patch size and TM rely on this area. Methods to extract human face from an image have been published in [2], [5], [16], [14], [12], [15] using color information, and [8], [1] using statistical learning approach. An overall survey for existing methods has also been done and published by Yang et al.[17]. 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 three techniques exist: (1)prepare a human-skin color model and calculate the distance from a pixel in local image to the distribution model using Euclidean, Mahalanobis[2][3] distance, etc., (2)fix a range value as the parameters for each domain in a color space through experiment[12], and (3)prepare a Look Up Table (LUT) for speeding up the computation time.

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.[14]. They

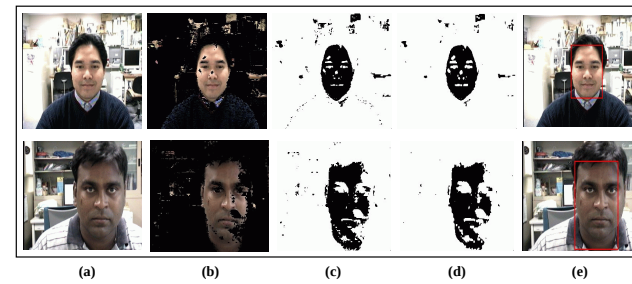

Figure 1. Detection of face region:(a)input image, (b)skin-color pixels, (c)after being binarized and region segmentation, (d)after background dilation, and (e)the face region is shown in rectangle.

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 be used as well. Readers interested in color-segmentation are referred to the work reported by Sigal et al. in [10] for the latest published material. 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 fixed the parameters for each domain as follows: $0.37 \leq \mathrm{r} \leq 0.47$ and $0.28 \leq \mathrm{g} \leq 0.35$. In Figure 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 skin 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 Figure 1(c) and (d), respectively. Then, we compute the vertical and horizontal projections using image in (d) to estimate the location of the face vertically and horizontally(Figure 1(e)).

3. Adaptive Face Model

The face model in our work is defined as a model consists of regions known as right eye region(RER), left eye region(LER) and mouth region(MOR)(shown in Figure 2). Interested features, which are right eye(REF), left eye(LEF) and mouth (MOF) will be searched within these three regions using TM. Each region in face model is computed by verifying two values, center of vertical and horizontal integral projections. The former divides the face region horizontally (*face horizontal center*(F_{hc})), while the latter divides the face region vertically (*face vertical center*(F_{vc})). However, only F_{hc} is used during tracking. Due to this difference, we associate two types of face model in EMoTracker. The definitions of each of them are given below:

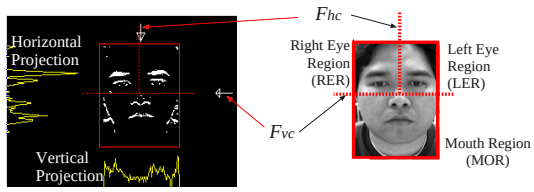

Figure 2. Analysis of vertical and horizontal integral projections to create face model.

- **Rigid face model** – this model is used only during initialization, which occurs either when EMoTracker 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 during tracking. Since searching for MOF is always done below both eyes using temporal information from previous frame, we skip computing F_{vc}. Thus, only RER and LER are known using this model. When the position of both pupils are known at time t, the area between both pupils at this time t will be utilized to define F_{hc} in the subsequent frame, $t+1$. Using F_{hc}, searching for each of corresponding facial component can be kept only within high possibility area.

Utilizing only rigid face model along the tracking process is insufficient in this framework. As values of F_{hc} and F_{vc} are computed from horizontal edge image [13], noise such as edges between hair and forehead, ears and back neck, influence the projections. As a result, false-positive F_{hc} is given which consequently effects the results. Figure 3 illustrates these problems. The red line shows F_{hc}. Adaptive face model shows correct F_{hc} even subject pans deeper to the right. We have also considered the area to perform the vertical projection so that noise can be reduced. As our tracking results provide size of eyes areas as well, we utilize the area shown in Figure 4 top-left image to determine F_{hc} using linear discriminant analysis[1].

4. Energy Minimization Criterion

In TM technique, a matched area in the image is given by the maximum correlation results without taking into consideration some other additional factors. However, in this tracking framework, while face moves, shades and shapes of facial features change with regards to face movement. This creates possibilities for the tracker to select incorrect area surrounding the target. An example is shown in Figure 5. To solve this problem, we propose a method that utilizes face geometrical constraints in addition to the correlation results. Matched areas are defined as the areas that have high correlation values and placed at a certain face geometrical constraint. These areas (referred to as "candidates") are also considered as the candidates that possess *energy minimization criterion*. Using TM, we list out n possible candidates by referring to their correlation values. They are arranged in descending order. Eyes candidates that minimize energy between eyes, E_{eyes} and mouth candidate that minimizes energy between eyes and mouth, E_{mouth} will be considered as the correct areas.

[1]Linear discriminant analysis is used for various applications. One of the example is Ohtsu's method [7], which is used for threshold selection.

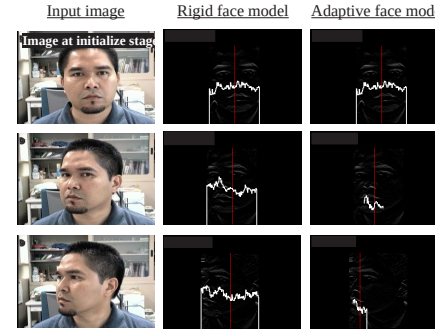

Figure 3. Vertical integral projection using different face model. Adaptive face model shown to be more reliable than rigid face model in handling variant face pose.

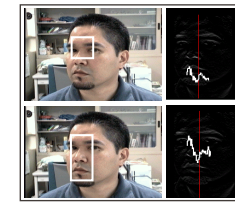

Figure 4. Vertical integral projection results from two different areas shown in rectangles. Images on left show the selected area, while images on right show the actual projection results.

4.1. Feature Selective

To increase reliability during TM, we propose *feature selective(FS)* technique. Using this technique, we seek the feature vector that best represents each corresponding facial component in face model regions. As a result, EMoTracker can automatically choose a suitable template depending on the appearance of the facial components. For example, a horizontal edges template would be suitable for a face with long front hair instead of intensity template, as we illustrate in Figure 6. Using templates made from a single feature vector for all subjects or for all facial components is unfortunately unreliable. For this purpose, we prepare two different feature vector templates, intensity and horizontal edges templates (made from haar horizontal transformed image).

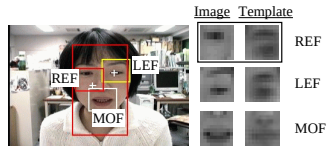

Figure 5. False-positive examples for REF during typical template matching(n=1) using intensity templates. On the right, comparison between matched areas and templates are shown.

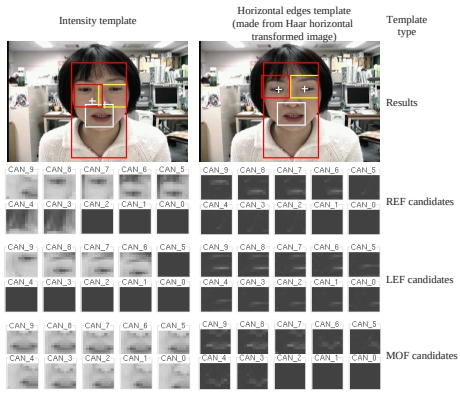

Figure 6. Results based on EMC (n=10) with a single feature vector template. An appropriate template type for each corresponding facial component is desired when multiple face type is considered.

4.2. Eyes Selection

Criteria to minimize the energy between both eyes are shown in these following items:

1. *correlation value* – we set the higher the correlation value is, the lower would the energy be,
2. *horizontal distance* between both eyes – a minimum horizontal distance between both candidates produces the minimum energy,
3. *vertical distance* between both eyes – a minimum vertical distance between both candidates produces the minimum energy.

The last two show that eyes are always near to each other in any face situation. Since TM for eyes are done independently for each eye, our tracker may handle detection of both eyes even the face tilts(eyes are placed diagonally) a little.

$$\begin{aligned} P_{E_w} &= (0.5) * F_w, \\ P_{E_h} &= (0.8) * P_{E_w}. \end{aligned} \tag{1}$$

$$\begin{aligned} E_{e_x}(i,j) &= \|\Delta x_{e(i,j)}\|/P_{E_w}, \\ E_{e_y}(i,j) &= \|\Delta y_{e(i,j)}\|/P_{E_h}, \\ E_{RC}(i) &= (1.0 - REyeCorrelation_i), \\ E_{LC}(j) &= (1.0 - LEyeCorrelation_j). \end{aligned} \tag{2}$$

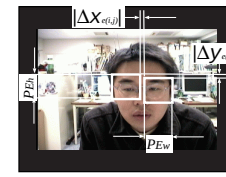

Figure 7. Parameters derived from face geometrical constraints to compute corresponding energies for selecting correct eyes.

$$\begin{aligned} \mathrm{E}_{\mathrm{eyes}}(i,j) &= \lambda_{e1}E_{e_x}(i,j) + \lambda_{e2}E_{e_y}(i,j) \\ &\quad +\lambda_{e3}E_{RC}(i) + \lambda_{e4}E_{LC}(j). \end{aligned} \tag{3}$$

Equation (2) shows how each parameter is normalized and equation to determine the minimum of total energy, $E_{eyes}(i,j)$ is defined in Equation (3). $E_{e_x}(i,j)$ and $E_{e_y}(i,j)$ are the horizontal and vertical energies between both eyes candidates (Figure 7). These energies are normalized with the patch width and height respectively, to overcome scaling problems. (i,j) is the index for right and left eye candidates, respectively. $\lambda_{e1} \sim \lambda_{e4}$ are the weighting parameters for each respective energy. In our work, $\lambda_{e1} \sim \lambda_{e4}$ are set to 1.0. In Figure 6, we show the results based on EMC using 10 candidates with a single feature vector template (for all interested facial components). Here, higher numbered candidates demonstrate higher similarity results while the black regions are actually showing that no candidate is listed. This happens because only candidates with correlation values higher than a fixed threshold value are considered.

Besides computing the energies, it is necessary to compute the patch size. In our work, we compute the eyes patch size using the equations defined in Equation (1). The computation is done with reference to the detected face width, F_w, thus, can be performed automatically along the tracking process. Face height is neglected when computing the patch height due to this value is an inconsistent value. Area where neck appears influent this value very much. P_{E_w}, P_{E_h} and F_w represent patch width, patch height and face width, respectively.

4.3. Mouth Selection

Selection for mouth is done after both eyes have been selected. Criteria to minimize the energy between eyes and mouth are shown in these following items:

1. *correlation value* – analogous to Section 4.2, we make the higher the correlation value is, the lower would the energy be,
2. *vertical distance* between mouth and eyes – a minimum vertical distance that complies with some restrictions shown in Equation (6) between lower side of the

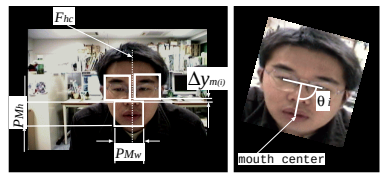

Figure 8. Parameters derived from face geometrical constraints to compute corresponding energies to select correct mouth. Left image shows how energy is computed relative to vertical distance, and right image shows definition of θ_i ($\Theta_i = 0.0$ if $\theta_i = 90.0$)

selected eyes to the upper side of mouth area produce the minimum energy,

3. *angle* – an angle formed by the intersection of two lines, *i.e.* line connecting both eyes and line connecting mouth and center of both eyes. As mouth is positioned perpendicular to center of the line connecting both eyes, we define the energy for this angle value as 0.0 when this assumption is complied, as shown in Equation (7).

Equation (8) defines the total of minimum energy to select correct mouth among the candidates. A few restrictions to fulfill mouth placement on face have also been associated. We define these in Equation (6) and Equation (7). In Figure 8, we illustrate corresponding parameters to minimize the total energy. (i) is the candidates index, $\lambda_{m1} \sim \lambda_{m3}$ are the weighting parameters for each energy. In the experiment, we set $\lambda_{m1} \sim \lambda_{m3}$ to 1.0. For mouth selection, we compute different patch size as shown in Equation (4). P_{M_w} and P_{M_h} represent mouth patch width and height, respectively.

$$\begin{array}{rcl} P_{M_w} & = & (1.2) * F_w, \\ P_{M_h} & = & (0.8) * P_{M_w}. \end{array} \tag{4}$$

$$\begin{array}{rcl} E_{m_y}(i) & = & \Delta y_{m(i)}/P_{M_h}, \\ E_{m_\theta}(i) & = & \Theta_i, \\ E_{MC}(i) & = & (1.0 - MouthCorrelation_i). \end{array} \tag{5}$$

$$\Delta\mathcal{M}_i = \begin{cases} 0.5 - E_{m_y}(i), & (E_{m_y}(i) < 0.0), \\ 0.0, & (0.0 \leq E_{m_y}(i) < 0.5), \\ 0.5 + E_{m_y}(i), & (0.5 \leq E_{m_y}(i)). \end{cases} \tag{6}$$

$$\Theta_i = \begin{cases} 90.0 - \theta_i, & (\theta_i <= 90.0), \\ \theta_i - 90.0, & (\theta_i > 90.0). \end{cases} \tag{7}$$

$$\mathrm{E}_{\mathrm{mouth}}(i) = \lambda_{m1}\Delta\mathcal{M}_i + \lambda_{m2}E_{m_\theta}(i) + \lambda_{m3}E_{MC}(i). \tag{8}$$

4.4. Pupils Detection

We detect pupils using TM. As the templates for these facial components look similar to eye brows, we search for these components within two third lower area of detected eyes. The darkest pixels within detected areas are assumed as the pupil centers. During tracking, searching for the darkest pixel is performed within previously detected area.

5. Experiments and Results

We have divided our experiment into two parts. First part is to evaluate the performance of proposed method in detection accuracy as well as the optimal candidates quantity. Typical TM technique is also performed in this part. In the second part, we evaluate the performance of proposed method in tracking manner. Results that are derived from the first experiment are utilized in this experiment. We also incorporate an adaptive face model in this experiment too.

For the test data, we prepared six color video sequences taken from different subjects. These subjects consist of three face types, *i.e.* normal face(NF), face with long front hair(LFH) and face with spectacles(FWS). Each of them was asked to move their face towards right and left with a slight pan, as well as forward and backward. Tracking precision was computed based on how many hits achieved compared to frame quantity. Here, a hit for *REF and LEF* means eye brows and eyes are included in the detected correct areas, a hit for *MOF* means lips are included in the detected correct area, and a hit for *each pupil* means the cross for corresponding pupil is made on the correct pupil.

Results for the first part of the experiment are shown in Table 1. Obviously, using typical template matching approach yields the worst result among all. Although it shows improvement when FS is utilized, using additional criteria that we have proposed has shown superior results above all observations. Failures are mostly caused by the failure of detecting a good face region and background noise, which caused by face pans to right and left more than approximately $30°$. From experiment, using more than 10 candidates with FS has shown a saturation state, therefore, we employ 10 candidates for detection purpose in our work. On the other hand, while 10 candidates is found to be the optimal candidates for detection purpose, we examine the optimal candidates for tracking purpose in the second experiment.

Table 1. Averaged detection precision for typical template matching (n=1) and proposed method.

Methods	Grayscale(%)	Haar(%)	FS(%)
TM(n=1)	87.5	88.9	89.1
EMC(n=10)	90.1	94.8	97.0

Table 2. Tracking results using adaptive face model in the second experiment.

Candidates	1	3	5	10
Ave. Tracking Precision(%)	95.7	98.3	98.4	97.9
Processing Time(msec)	2.51	2.73	2.85	2.98

As usually performed in tracking framework, we utilize previously detected positions to search for corresponding components in current frame. We initially start with 10 candidates and repeat the same experiment using 5, 3 and 1 candidate(s). Surprisingly, using less candidates with adaptive face model has shown an increment in tracking accuracy. But however, when only 1 candidate is utilized, the worst result is observed. This is identical to using TM in tracking framework. In our work, our interested facial components are not small areas. Although searching at previously detected position leads to high possibility detection, yet the problem of appearance changes is remained unsolved. Above all this, the processing time for each of 3 and 5 candidates is only 2.73 msec and 2.85 msec, respectively, which ensure us that our proposed method has a high reliability in tracking precision and feasible for real-time applications. (For reference, our experiments were performed on a Celeron 2.2GHz CPU with 512 MByte of memory. The OS is FreeBSD 4.7 release.)

6. Conclusion

We have introduced a novel method to detect and track facial components in real-time. This method has been embedded into a tracker given by a name, *EMoTracker*. Our method determines the correct components based on energy minimization criterion. Instead of utilizing the similarity or correlation in template matching independently, we take into consideration some face geometrical constraints about the placement of corresponding facial components on face to compute related energies. Candidates that minimize these energies are considered as the correct areas. Additionally, we also proposed feature selective technique, which showed empirically to improve results using typical template matching. On top of that, we have further improved our method by incorporated adaptive face model into EMoTracker. More than 98% of tracking precision has been achieved in our work that requires approximately about 2.7$\sim$2.9 msec of processing time. In the near future, we plan to improve the performance by adjusting the template size with respect to face pose.

References

[1] A. J. Colmenarez. *Facial Analysis From Continuous Video With Application to Human-Computer Interface*. PhD thesis, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1999.

[2] 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*, 1793:197–206, April 2000.

[3] 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)*, volume 2, pages 300–305, 1996.

[4] 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*, volume 2, pages 1525–1530, 1997.

[5] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, May 2002.

[6] 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)*, volume II, pages 830–835, Oct, 12-15 1999.

[7] N. Ohtsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-9(1):62–66, 1979.

[8] 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.

[9] K. Sakaue and K. Yamamoto. Active net model and its application to region extraction. *The Journal of the Institute of Television Engineers of Japan*, 45(10):1155–1163, 1991. in Japanese.

[10] L. Sigal, S. Sclaroff, and C. Athitsos. Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):862–877, July 2004.

[11] P. Smith, M. Shah, and N. da Vitoria Lobo. Monitoring head/eye motion for driver alertness with one camera. In *Fifteenth IEEE International Conference on Pattern Recognition, Barcelona, Spain*, September 3-8 2000.

[12] K. Sobottka and I. Pitas. Extraction of facial regions and features using color and shape information. In *International Conference on Pattern Recognition (ICPR)*, volume III, pages C421–C425, 25–29 August 1996.

[13] 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 Proceedings*, 2003.

[14] 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 Sensing Via Image Information*, pages 457–462, 2002.

[15] J. Yang, W. Lu, and A.Waibel. Skin-color modeling and adaptation. Technical Report CMU-CS-97-146, Carnegie Mellon University, May 1997.

[16] M.-H. Yang and N. Ahuja. Detecting human faces in color images. In *Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP 98)*, volume 1, pages 127–130, October, 1998.

[17] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.