Lecture Notes in Artificial Intelligence, vol. 1793, pp. 197-206, April 2000, Springer-Verlag press http://www.springer.de/comp/lncs/index.html (MICAI-2000, Acapulco)

Detection and Tracking of Facial Features in Video Sequences

Rogério Schmidt Feris
Teófilo Emídio de Campos
Roberto Marcondes Cesar Junior

{rferis, teo, cesar} @ime.usp.br
CreatiVision - Creative Vision Research Group
Department of Computer Science
DCC-IME-USP, University of São Paulo
Rua do Matão, 1010, São Paulo, SP, 05508-900, Brazil
Phone: +55 (011) 818 6235, Fax: +55 (011) 818 6134

Abstract. This work presents a real time system for detection and tracking of facial features in video sequences. Such system may be used in visual communication applications, such as teleconferencing, virtual reality, intelligent interfaces, human-machine interaction, surveillance, etc. We have used a statistical skin-color model to segment face-candidate regions in the image. The presence or absence of a face in each region is verified by means of an eye detector, based on an efficient template matching scheme . Once a face is detected, the pupils, nostrils and lip corners are located and these facial features are tracked in the image sequence, performing real time processing.

1 Introduction

The detection and tracking of faces and facial features in video sequences is a fundamental and challenging problem in computer vision. This research area has many applications in face identification systems, model-based coding, gaze detection, human-computer interaction, teleconferencing, etc.

In this paper, we present a real time system that performs facial features detection and tracking in image sequences. We are interested in developing intelligent interfaces and this system is the first step to give computers the ability to look at people, which may be a good way to improve human-computer interaction.

We have used a color-based approach to detect the presence of a human face in an image. A Gaussian distribution model was generated by means of a supervised training, using a set of skin-color regions, obtained from a color face database.

Basically, the input image is compared with the generated model in order to identify face candidates, i.e., skin-color regions. The presence or absence of a face in each region is verified by using an eye detector based on an efficient template matching scheme. There is biological evidence that eyes play the most important role in human face detection [1].

Once a face is detected, the pupils, nostrils and lip corners are located and tracked in the video sequence. The system is very fast and robust to small face pose and rotation variation.

The remainder of this paper is organized as follows. Section two reviews some techniques related to our work. Section three describes the statistical skin-color model

and section four explains how it was used to segment face-candidate regions. Section five describes the eye detector used to perform the face verification process. In section six, the detection and tracking of facial features is described. The obtained results are showed in section seven. Finally, in section eight follows a conclusion and we address further research directions.

2 Related Work

Automatic detection of facial features in video sequences is performed, in general, after a face is detected in a frame. Many approaches have been proposed to detect faces in static scenes, such as Principal Component Analysis [1,2], clustering [3] and neural nets [4]. These techniques obtained good detection / false alarm rates, but they are computational expensive and hardly achieve real-time performance.

Liyanage Silva et. al. [5] proposed a method, which they called edge pixel counting, to detect and track facial features in video sequences. This method is based on the fact that a higher edge concentration is observed in the vicinity of facial features, while slightly outside of such features a less edge concentration is observed. In spite of its simplicity, the method fails in the presence of cluttered backgrounds, glasses, hat, etc.

Recently, several color-based systems have been proposed to perform face detection and tracking in image sequences. Processing color is much faster than processing other facial features and, furthermore, color is orientation invariant under certain lighting conditions. This property makes motion estimation much easier since only a translation model is needed.

The work of Jie Yang and Alex Waibel [6] presents a statistical skin-color model, adaptable to different people and different lighting conditions. This work was extended to detect and track faces in real time [7], achieving a rate of 30+frames/second using a HP-9000 workstation with a framegrabber and a Canon VC-C1 camera.

Nuria Oliver et. al. [8] and Trevor Darrell et. al. [9] also obtained good results using skin-color to detect and track people in real time. Stiefelhagen and Yang [10] presented a method similar to ours, which detects and tracks facial features in image sequences. The main differences lie in face verification process. They have used iterative thresholding to search the pupils whereas we have adopted an efficient template matching scheme to detect eyes.

3 Skin-color Model

The statistical skin-color model [6] is generated by means of a supervised training, using a set of skin-color regions, obtained from a color face database, with 40 face images. Such images were obtained from people of different races, ages and gender, with varying illumination conditions. Figure 1 illustrates the training process, in which a skin-color region is selected and its RGB representation is stored. It was verified, using training data, that skin-colors are clustered in color space, as illustrated in Figure 2.

It is common to think that different people have skin-colors which differ significantly from each other due to the existence of different races. However, what really occurs is a larger difference in brightness / intensity and not in color [6]. Thus,

it is possible to reduce the variance of the skin-color cluster through intensity normalization:

$$a = \frac{R}{R+G+B} \tag{1}$$

$$b = \frac{G}{R+G+B} \tag{2}$$

The colors (a,b) are known as chromatic or “pure” colors. Figure 3 illustrates the skin-color cluster in chromatic space.

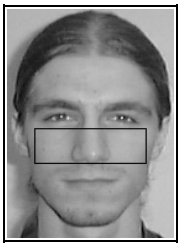

Fig. 1. Selected skin-color region

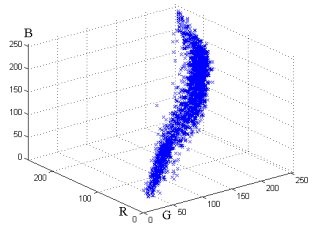

Fig. 2. Cluster in color space

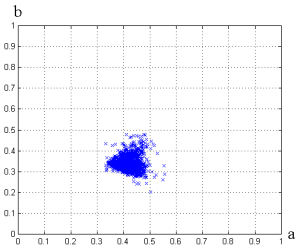

Fig. 3. Cluster in chromatic space

According to [6], the skin-color distribution in chromatic space can be approximated by a Gaussian model $N(m,\Sigma^2)$, where $m=(a_{avg},b_{avg})$ and:

$$a_{avg} = \frac{1}{N} \sum_{i=1}^{N} a_i \tag{3}$$

$$b_{avg} = \frac{1}{N} \sum_{i=1}^{N} b_i \tag{4}$$

$$\Sigma = \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ba} & \sigma_{bb} \end{bmatrix} \tag{5}$$

The Gaussian model, which was generated from training data, is illustrated in Figure 4.

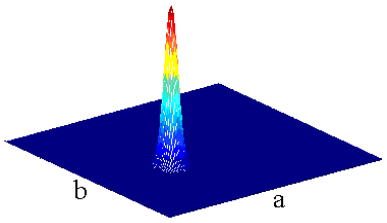

Fig. 4. Gaussian Model

4 Image Segmentation

The segmentation process consists in identifying skin-color regions in the input image (the first frame of the video sequence) based on the generated model.

Initially, all pixels in input image are converted to the chromatic space. Using the Gaussian model, a gray-level image is generated and the intensity of each pixel in this new image represents its probability to have skin-color. Figure 5 illustrates this process.

Since skin-color pixels present a stronger response in the gray-level image, we can identify such regions by a thresholding process. The threshold value was chosen empirically.

In general, due to noise and distortions in input image, the result of thresholding may generate non-contiguous skin-color regions. Furthermore, after this operation, spurious pixels may appear in the binary image. In order to solve these problems, first we have applied a morphological closing operator to obtain skin-color blobs. Subsequently, a median filter was used to eliminate spurious pixels. Figure 6 shows the obtained binary image after thresholding (left illustration) and the result of morphological closing operation followed by a median filtering (right illustration).

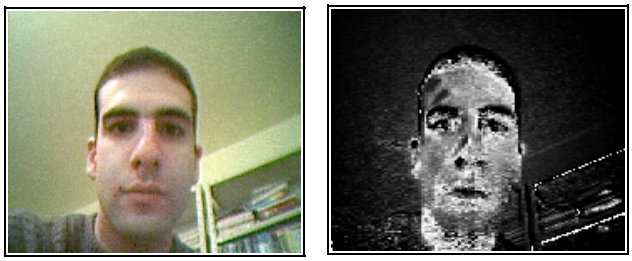

Fig. 5. Skin-color identification

Fig. 6. Thresholding and morphological closing followed by median filtering

Boundaries of skin-color regions are determined using a region growing algorithm in the binary image. Regions with size less than 1% of image size are eliminated. Furthermore, structural aspects are considered so that regions that do not present the face structure are removed.

In the next section, we will show a method to decide whether each detected skin-color region corresponds or not to a face.

5 Verification Process

The main goal in identifying skin-color regions in the input image is to reduce the search space for faces. However, it is important to note that not all detected regions contain faces. Some correspond to parts of human body, while other correspond to objects with colors similar to those of skin. Thus, it is necessary to verify the presence or absence of a face in each detected region.

To accomplish this task, we have used an eye detector based on an efficient template matching scheme. It is worth saying that there is biological evidence that eyes play the most important role in human face detection [1]. The eye template used is illustrated in Figure 7. Its resolution is 57x15 with 256 gray levels.

Fig. 7. Eye template

The eye detector procedure first converts the skin-color region to a gray-level image, in order to perform template matching. The most important property in our

technique is that the eye template is resized according to the skin region size. In general, existing methods use several templates with different scales to accomplish the detection. For instance, the work of Brunnelli and Poggio [11] uses five templates with scales 0.7, 0.85, 1, 1.5 and 1.35 to detect eyes in an image. Although these methods have obtained good results, they are computational expensive and not suitable to real time processing.

Our method estimates the person eye size, resizing the eye template. Thus, the template matching is performed only once, using the upper portion of the detected skin-color region. Basically, correlation coefficients ρ are computed in different positions within the skin region:

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma(x)\sigma(y)} \tag{6}$$

where x is the eye template and y is the image patch in a specific position within the skin region. The function *cov* in the equation above returns the covariance between x and y. The correlation coefficient ρ assume values ranging from -1 to 1.

A threshold value, which was determined empirically, was used to decide eye detection. When eyes are found within the skin-color region, we report the presence of a face. The system then detects some facial features and proceeds in the tracking module, as described in the next section.

It is important to note that the above strategy detects the first frame in which the face appears in frontal view.

6 Detection and Tracking of Facial Features

Once a face is detected, the system proceeds with the search for pupils, nostrils and lip corners. After, these facial features are tracked in the video sequence. The approach used is similar to the work of Stiefelhagen and Yang [10].

6.1 Searching the Pupils

Assuming a frontal view of the face initially, we locate the pupils by looking for two dark pixels that satisfy certain geometric constraints and lie within the eye region detected by template matching.

6.2 Searching the Lip Corners

First, the approximate positions of lip corners are predicted, using the position of the face and the pupils. A small region around those points is then extracted and a Sobel horizontal edge detector is applied. The approximate horizontal boundaries of the lips are determined by first computing the vertical integral projection P_v of this horizontal edge image:

$$P_V(y) = \sum_{x=1}^{H} E_h(x, y), \quad 0 \le y \le W \tag{7}$$

where $E_h(x,y)$ is the intensity function of the horizontal edge image, and W and H are the width and height of the search region, respectively.

The approximate left and right boundaries of the lips are located where P_v exceeds a certain threshold t or respectively falls below that threshold. We choose t to be the average of the projection P_v. The vertical position of the left and right lip corners are simply located by searching for the darkest pixel along the columns at the left and right estimated boundaries of the lips in the search region.

6.3 Searching the Nostrils

Similar to searching the pupils, the nostrils are located by searching for two dark regions, which satisfy certain geometric constraints. The search region is restricted to an area below the pupils and above the lips.

Figure 7 illustrates the detection of facial features.

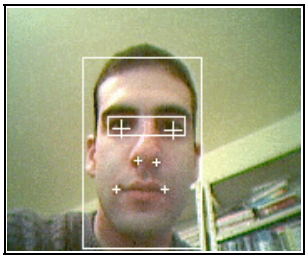

Fig. 7. Detection of facial features

6.4 Tracking

For tracking pupils and nostrils in the video sequence, simple darkest pixel finding in the search windows around the last positions is used. In order to track lip corners, the same detection process is performed in a small region around the last lip corners position.

7 Obtained Results

The proposed real time facial feature tracker obtained good results in different video sequences. Initially, a color-based frontal view face detection is performed. This procedure fails, in general, when the eye template scale is incorrectly estimated. This case occurs when the skin region contains, beyond a face, other human body parts, such as a hand occluding the face, or other objects. We have also noted that faces may be missed when they are too small or under strong varying illumination.

The face detection procedure has proven to be robust in identifying skin-colors of people with different races, ages and gender. It is worth saying that dark-skin regions had smaller skin-color probabilities, since most persons in the face database had light skin.

Once a face is detected, its facial features are located and tracked in the image sequence. If there is more than one face in the first frame, only the bigger face is

considered.

During the tracking process, we verified that the system fails when a large movement is performed by the person. Furthermore, due to varying illumination, some features may be missed in the sequence. On the other hand, the tracking module showed to be robust to small face pose and rotation variations.

Figure 8 shows the tracking process in some frames of a image sequence, which can be observed in *http://www.ime.usp.br/~cesar/creativision*.

Fig. 8. Tracking Process

8 Conclusions

This paper described a real time system for detection and tracking of facial features in video sequences. A statistical skin-color model was used to detect face-candidate regions and an efficient template matching scheme was applied to report the presence or absence of a face in these regions. Given a detected face, pupils, nostrils and lip corners are located and tracked in the image sequence. The system may be used in different applications, such as teleconferencing and human-computer interaction.

Future work includes pose estimation and 3D head tracking by using the positions of pupils, nostrils and lip corners. Furthermore, we intend to improve the system robustness developing a recovery module and considering motion to increase the accuracy of detection and tracking of facial features.

Acknowledgments

Roberto M. Cesar Jr. is grateful to FAPESP for the financial support (98/07722-0), as well as to CNPq (300722/98-2). Rogério Feris and Teófilo Campos are grateful to FAPESP (99/01487-1 and 99/01488-8).

We are grateful to Carlos Hitoshi Morimoto (IME-USP) for useful discussions.

References

[1] B. Moghaddam and A. Pentland. “Face Recognition using View-Based and Modular Eigenspaces”. Proc. of Automatic Systems for the Identification and Inspection of Humans, SPIE vol. 2277, July 1994.

[2] B. Moghaddam and A. Pentland. “Probabilistic Visual Learning for Object Detection”. In Fifth International Conference on Computer Vision, pp. 786-793, Cambridge, Massachusetts, June 1995

[3] K. Sung and T. Poggio. “Example-base Learning for View-based Human Face Detection” C.B.C.L. Paper no. 112, MIT, 1994.

[4] H. Rowley, S. Baluja and T. Kanade. “Neural Network-based Face Detection”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, January 1998.

[5] L. Silva, K. Aizawa and M. Hatori. “Detection and Tracking of Facial Features”. Proc. of SPIE Visual Communications and Image Processing, Taiwan. May, 1995.

[6] J. Yang, W. Lu and A. Waibel. “Skin-Color Modeling and Adaptation”. CMU CS Technical Report, CMU-CS-97-146. May, 1997.

[7] J. Yang and A. Waibel. “A Real-time Face Tracker”. Proc. of the Third IEEE Workshop on Applications of Computer Vision, pp. 142-147, Sarasota, Florida, 1996.

[8] N. Oliver, A. Pentland and F. Berard. “LAFTER: Lips and Face Real-time Tracker with Facial Expression Recognition”. Proc. of Computer Vision and Pattern Recognition Conference, 1997.

[9] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland. “Pfinder: Real-time Tracking of the Human Body”. Proc. SPIE, vol. 2615, pp. 89-98, 1996.

[10] R. Stiefelhagen and J. Yang. “Gaze Tracking for Multimodal Human Computer Interaction”. University of Karlsruhe, 1996. Available at http://werner.ira.uka.de/ISL.multimodal.publications.html.

[11] R. Brunelli and T. Poggio. “Face Recognition: Features versus Templates”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042-1052. October 1993.