

Automatic Face and Facial Features Initialization for Robust and Accurate Tracking

Murad Al Haj, Javier Orozco, Jordi González, and Juan J. Villanueva

*Centre de Visio per Computador, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
malhaj@cvc.uab.es*

Abstract

Face detection and tracking, through image sequences, are primary steps in many applications such as video surveillance, human computer interface, and expression analysis. Many currently existing techniques don't perform well due to pose variations, appearance changes, illumination changes, complex backgrounds, and inaccurate initialization. The last short coming, which is the difficulty to initialize motion regions, is a problem facing any tracker. In this paper, we present an automatic and robust face detection and tracking system for color image sequences. Face detection is done using skin color segmentation and connected components analysis. Later, facial features are detected by active shape models and a face mesh is initialized. Finally, the tracking is done by active appearance models. Experimental detection and tracking results on a pose varying face video are given.

1. Introduction

Detecting and tracking faces in images are essential in many applications of computer vision. Of these applications, facial expression analysis and human computer interaction have attracted much interest lately. The main challenges facing such a system, specially in an indoor scenario, include:

- (1) Appearance Changes which are mainly due to the variation of the camera-face pose, causing the face to be frontal or profile or in between.
- (2) Illumination Changes which are abrupt in the case of indoor scenes, unlike the gradual changes that take place in outdoor scenes.
- (3) Complex backgrounds, especially in indoor scenes, where many objects are present such as chairs, tables, and doors; and these objects can be moving which worsens the problem.

Many existing techniques try to solve these problems in order to achieve robust detection and tracking. Starting with the detection problem, a survey for face detection is presented, [10], where the different methods are classified into four, sometimes overlapping, categories: knowledge-based methods, feature invariant approaches, template matching methods, and appearance based methods. Viola and Jones, [8], have introduced a rapid and robust object detection method using Haar-like features and Adaboost and they applied it to face detection. This method has been gaining popularity among researchers; however, the problem with this method is that it is sensitive to the face pose. Some modifications have been proposed to make this approach more robust for multi-view [6], but the problem of fast multi-view face detection is still an open field [9].

Experience proved that skin is an effective and robust cue for multi-view face detection. Color is highly invariant to geometric variations of the face and it allows fast processing. Therefore, in this paper we will use skin color segmentation as a first step towards face detection.

Color-based tracking methods have many advantages, they are robust against appearance changes and complex backgrounds. Two examples of widely used color-tracking methods are Meanshift [3], and Camshift [2]. These methods have a shortcoming due to their inability to initialize motion regions, which can be solved by the detection process. However, the inherited nature of these methods to track complete regions makes them successful in tracking the whole face, but inapplicable when it comes to tracking facial features.

In this paper, a complete automatic face detection and tracking system is presented, see Fig. 1. The method can be summarized into 3 main parts: Face Detection, Facial Features Detection followed by Mesh Initialization, and Face Tracking. The paper is organized according to these parts, Section 2 introduces our proposed face detection method. Section 3 discusses the

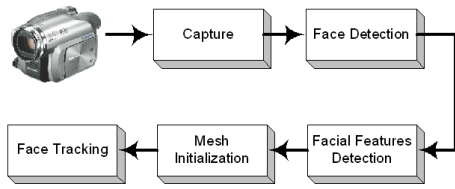


Figure 1: An overview of the system

facial features detection and mesh initialization, while section 4 presents the tracking method. In section 5, experimental results are shown. Concluding remarks are discussed in section 6.

2. Face Detection

Our proposed face detection method is based on skin color. It starts by segmenting the image into regions that contain possible face candidates, while those that do not contain a face object are dropped. This segmentation helps accelerate the detection process. Later, connected components are analyzed and some primitive shape features of the human face are used to decide which region is a face and which is not.

2.1. Skin Color Model

In this section, we present a classifier for skin color based in the RGB colorspace. The proposed algorithm is as follows:

$$\begin{aligned}
 (R, G, B) \text{ is classified as skin if:} \\
 20 < R-G < 90 \\
 R > 75 \\
 R/G < 2.5
 \end{aligned}$$

The advantages of this method are its simplicity and computational efficiency since no transformation is needed to go to another colorspace. In our experiments, around 800,000 skin pixels were taken from 64 different images of people with different ethnicities and under various illumination conditions. The distribution of R-G is shown in Fig. 2.(a), while that of R is shown in Fig. 2.(b), and that of R/G is shown in Fig. 2.(c). Our experiments revealed that 94.6% of the skin pixels have their R-G values between 20 and 90, see Fig. 2.(a), which supports the observation in [1]. Also, we have noted that 96.9% of the skin pixels have their R value greater than 75, see Fig. 2.(b), and 98.7% of them have their R/G values less than 2.5, see Fig. 2.(c).

2.2. Component Analysis

After the segmentation process, a sobel edge detector is used to separate possible face candidates from any

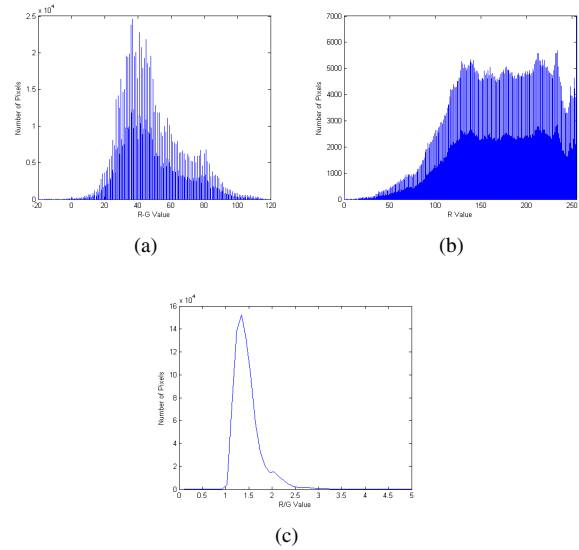


Figure 2: (a) The distribution of R-G values in skin pixels (b) The distribution of R values in skin pixels (c) The distribution of R/G in skin pixels

background object. After executing the edge detection, a dilation operation is applied to further separate the edges. The resulting image is then searched for connected components using the 8-neighbor pixels. Each of the connected components is later analyzed to decide whether it is a face or not. Simple shape features are used in the classification process. Each of these features, or cues, can be thought of as a weak classifier. The cascade of these weak classifiers forms a strong classifier, as shown later in the experimental results. These features are:

1. Area: Components with small areas are dropped.
2. Bounding Box Proportions: Any region whose height is more than 1.8 times its width is removed.
3. Holes: Since not all the face is covered with skin, any region with no holes is deleted.
4. Centroid: The face is evenly distributed in the region where it is located. Therefore, the centroid of a face region should be found in a small window centered in the middle of the bounding box. Any region whose centroid is outside this window is dropped.
5. Extent: The extent of a blob is the area of this blob divided by the area of the bounding box. Our experiments revealed that the extent of a face is between 0.3 and 0.8. Thus, any region whose extent is not in this range is eliminated.



Figure 3: (a) Features Detection (b) Mesh Initialization

3. Facial Features Detection and Mesh Initialization

Once the face is detected, facial features are extracted using active shape models. Active Shape Models, proposed by Cootes et al. [4], were extended and applied to facial landmarks detection by Milborrow [7]. The main extensions done by Milborrow were using two landmark profiles instead of one, and stacking two active shape models in series. This process outputs 68 facial points corresponding to the main features of the face, as shown in Fig. 3.(a).

It should be noted that this process is only applied once a frontal face is detected, that is because the presence of a frontal face is necessary to initialize the tracking of the various facial features. Frontal faces can be distinguished from profile faces using the ratio of the bounding box width to its length.

Once the image coordinates of the facial features are determined, a 3D Candide face model is fitted, as shown in Fig. 3.(b). The Candide face model will be used in tracking the face and the movement of the eyebrows, lips, and eyelids. It is given by the three spatial coordinates of each vertex. The shape is described by the $(n \times i)$ matrix \mathbf{F} , where n is the number of vertices and i is their coordinates:

$$\mathbf{F}_n^i = \mathbf{f}_n^i + \mathbf{D}_n^{i,d} \vartheta_d + \mathbf{E}_n^{i,e} \gamma_e \quad (1)$$

where \mathbf{f} is the default configuration, \mathbf{D} encodes the biometry of each person, and \mathbf{E} handles the facial animation. We consider $d=16$ with the biometric parameters represented by the vector ϑ , and $e=7$ with the facial action parameters represented by the vector γ that encodes the position of the eyebrows, eyes, and lips.

The tracking vector \mathbf{q} is then extracted, which contains the head pose (three Eulers angles, scale value, and image coordinates) and the seven animation parameters. The tracking vector is given by:

$$\mathbf{q} = [\alpha, \gamma] = [\theta_x, \theta_y, \theta_z, s, t_x, t_y, \gamma_0, \dots, \gamma_6] \quad (2)$$

4. Face and Facial Features Tracking

Given an image sequence \mathbf{I}_t , depicting head motion and facial expressions, we model each face by constructing an appearance-based model, which projects the 3D mesh onto the input image for a specific configuration of the vector \mathbf{q}_t .

The tracking is based on the method proposed in [5], where in order to estimate the corresponding vector \mathbf{q}_t at each frame, a corresponding appearance model, \mathbf{A}_t , is constructed by applying a warping process, $\psi(\mathbf{I}_t, \mathbf{q}_t) \rightarrow \mathbf{A}_t(\mathbf{q}_t)$. Consequently, the appearance model depends on the vector \mathbf{q}_t and the animation parameters γ_t .

Each appearance model, \mathbf{A}_t , is assumed to follow a Gaussian distribution, $N(\mu_t, \sigma_t)$. Therefore, we can apply a time-efficient filtering technique to estimate the Gaussian parameters over time with respect to the previous estimations. All estimated appearances, $\hat{\mathbf{A}}$, are held under an exponential with an updating factor ω as follows:

$$\begin{aligned} \mu_{t+1} &= \omega \mu_t + (1 - \omega) \hat{\mathbf{A}}_t \\ \sigma_{t+1}^2 &= \omega \sigma_t^2 + (1 - \omega) (\hat{\mathbf{A}}_t - \mu_t)^2 \end{aligned} \quad (3)$$

An adaptive velocity model is adopted in order to estimate the vector \mathbf{q}_t . The current input image of \mathbf{I}_t , at a certain instance t , is registered with the current appearance model \mathbf{A}_t , which depends on the estimated vector $\hat{\mathbf{q}}$. The final estimation is obtained by minimizing the Mahalanobis distance between the estimated and the current average appearances. Here, the appearance parameters, μ and σ , are known, and the distance is minimized by an iterative first-order linear approximation and calculating the Jacobian matrix:

$$\begin{aligned} \mathbf{A}_t &\approx \hat{\mathbf{A}}_{t-1} + \frac{\partial(\mathbf{A}_t, \mathbf{q}_t)}{\partial \mathbf{q}_t} (\mathbf{q}_t - \hat{\mathbf{q}}_{t-1}) \\ \Delta \mathbf{q}_t &= \mathbf{q}_t - \hat{\mathbf{q}}_{t-1} = -\mathbf{J}_t^* [\psi(\mathbf{I}_t, \hat{\mathbf{q}}_{t-1}) - \mu_t] \end{aligned} \quad (4)$$

where \mathbf{J}_t^* is the pseudo inverse of the Jacobian matrix. Thus, we apply a gradient descent method by partial differences, which is able to accommodate appearance changes while achieving precise estimation.

5. Experimental Results

In this section, we will show the results of the detection process independently from those of tracking. The detection method was tested on different pictures for people from both genders, from various ethnicities with varying skin color, and under changing lighting conditions. We were able to achieve a high detection rate (over 85%). Since we are interested in detecting multi-view faces in video sequences we will show the results

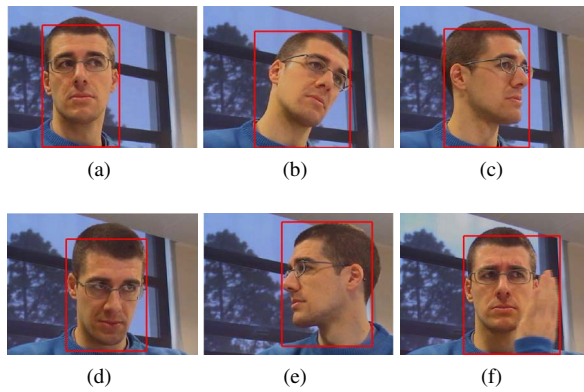


Figure 4: Detection Experimental Results

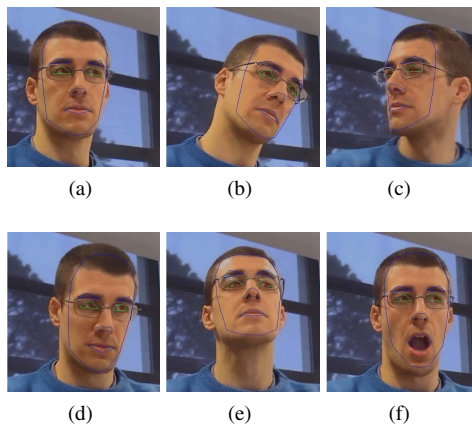


Figure 5: Tracking Experimental Results

on one sequence where the face has varying poses. We were able to correctly detect the face in almost all the images, a sample of the results is shown in Fig. 4. This method is successful in detecting multi-view faces however a disadvantage arises when the face is partially occluded by a hand, or other skin-colored object, where this object is detected as part of the face as shown in Fig. 4.(f). It should be remembered here that the detection is the initial step towards tracking; we tested the detection on the whole sequence only for demonstration purposes. Once a frontal face is detected, tracking is done using an active appearance model. We were able to correctly track the face and the different facial parts throughout the whole sequence with all the variations in pose. Some of the results are shown in Fig. 5.

6. Conclusions

In this paper a complete automatic system for detecting and tracking faces and facial features in video

sequences is presented. The advantages of this system are its low computational cost and robustness. Once a frontal face is detected, facial features are extracted, a mesh is initialized, and tracking is done through active appearance model. Future work will focus on using the tracking information for expression analysis.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDIVideo project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 MIPRCV CSD2007-00018. Jordi Gonzalez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

- [1] S. A. Al-Shehri. A simple and novel method for skin detection and face locating and tracking. In *APCHI*, pages 1–8, 2004.
- [2] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2):15, 1998.
- [3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the 16th Intl. Conf. on Pattern Recognition (ICPR'00)*, volume 2, pages 142–149, 2000.
- [4] T. Cootes, D. Cooper, C. Taylor, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan. 1995.
- [5] F. Dornaika, J. Orozco, and J. González. Combined head, lips, eyebrows, and eyelids tracking using adaptive appearance models. In *the 4th Intl. Conf. on Articulated Motion and Deformable Objects (AMDO'06)*.
- [6] C. Huang, H. Ai, B. Wu, and S. Lao. Boosting nested cascade detector for multi-view face detection. In *Proceedings of the 17th Intl. Conf. on Pattern Recognition (ICPR'04)*, volume 2, pages 415–418, Aug. 2004.
- [7] S. Milborrow. *Locating Facial Features with Active Shape Models*. Master's thesis. University of Cape Town (Department of Image Processing), 2007.
- [8] P. Viola and M. Jones. Robust real-time object detection. In *Proceedings of IEEE Workshop on Statistical and Computational Theories of Vision*, July 2001.
- [9] Y. Wang, Y. Liu, L. Tao, and G. Xu. Real-time multi-view face detection and pose estimation in video stream. In *Proceedings of the 18th Intl. Conf. on Pattern Recognition (ICPR'06)*, volume 4, pages 354–357, 2006.
- [10] J. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Tras. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):34–58, Jan. 2002.