Detection and Tracking Facial Features under Complex Background

ZHUANG Li   XU Guang-you   AI Hai-zhou   SONG Gang
Department of Computer Science and Technology, Tsinghua University
State Key Laboratory of Intelligent Technology and Systems

ABSTRACT

A “coarse-to-fine” facial feature detection and tracking system which is used under complex background is introduced in this paper. The system uses stereo cameras for video input. By stereovision technique, face is roughly and quickly segmented from complex background. Then, the multiple template matching method is applied to find the accurate face region from this rough segmentation. Facial organ candidates are extracted from the detected face region at a specific scale space called organ scale for Sobel filter. Finally, eyes, nose and mouth corners are detected. Techniques for checking and correcting errors in facial feature detection based on multiple cues are developed to make the algorithm more robust in facial feature detection and tracking in video sequence. Experiments on 189 video sequences demonstrate its effectiveness.

Keywords:  Face detection, facial feature detection and tracking, multi-scale analysis, face recognition

1. INTRODUCTION

Face feature detection is one of the cruxes in face recognition. The most difficult problem in face recognition is how to deal with the effect of illumination and pose changes. In a complex environment, the difference of the same person under different light and poses is sometimes more significant than that of different persons under the same light and pose. The incoherence of face data from different environment becomes a big obstacle in face recognition. J.Daugman [1] and Maxim A.Grudin [2] point out that: Current algorithm in face recognition is not robust enough with variable illumination and poses.

Current method in face recognition basically can be divided into two categories: feature-based and template-matching-based [3]. Both are based on facial feature detection. The feature-based approaches have more tolerance with illumination and poses: detect feature first, and then describe features. In template-based approaches, feature detection is needed in order to normalize raw data in geometry and illumination. All together, facial feature detection is a key point in improving the performance of face recognition. Many researchers [2][4][5][6] have regarded the robustness and correctness of feature detection as an important approach in face recognition.

Compared with static pictures, the videos have more information. It is easy to get affluent face samples of different poses from video sequence. Using video input can be very helpful to make face recognition into reality. In the context of video sequence, facial feature detection and tracking is one of the basic problems. This paper focuses on detection and tracking facial feature under complex background in video sequence.

Current algorithm of facial feature detection is not robust enough to light and pose changes, so its performance under complex background is poor. To this problem developing techniques for checking and correcting errors in facial feature detection can be very helpful. These techniques should basically depend on multiple cues related to facial feature detection that are complement with each other. We think that face detection and facial feature detection are closely related to each other: if face region can be detected reliably, facial feature could be detected based on the face model; contrariwise, facial feature can be used in checking the correctness of face detection, and thus boost the reliability of face detection. We combined these two aspects in our system. Face detection is used as the first step, using the global model to instruct local feature detection. Considering the face with little expression, we can regard faces under different poses as rigid bodies that follow rigid body motion rule. This is the basic assumption used in verifying the validity of detected features.

In this paper, we proposed a multi-cue based coarse-to-fine facial feature detection and tracking approach under complex background. In order to detect face region robustly, depth information in stereovision is combined with multiple template face detection algorithm. Eigen-eyes and multi-view facial feature checking method is used in verifying the correctness of facial feature. The system has the ability of checking errors and sometimes correcting errors.

2. BASIC ALGORITHM AND SYSTEM SETUP
The system setup is shown in Fig. 1 that uses two synchronized cameras for stereo video input. The two cameras are separated about 50cm. A person is about 3 meters away in front of the two cameras. A Pentium-III-966MHz PC with two Matrox Meteor II image grabbers is used as the platform. The control panel can control the orientation and the lens of the cameras to guarantee a suitable view field. We use stereo technique since it has been matured enough to be used for robust foreground segmentation that makes the system more adaptive to complex background environment.

The basic algorithm of facial feature detection and tracking is shown in Fig. 2. The whole process can be divided into two stages: detection and tracking. In the stage of detection, a whole face pattern is used to instruct local feature detection; the detection results of several frames are used to verify each other, establish a reference when it has the correct result. Then in the stage of tracking, the established correct reference is used to instruct next frame’s face detection, and then facial feature tracking.

![Fig. 1 System Setup](image)

![Fig. 2 Basic Algorithm](image)

3. FACIAL FEATURE DETECTION

3.1. ROUGH SEGMENTATION BY STEREO VISION

Supposing the target person stands in front of all other background objects, it will have the highest disparity in the disparity map. As shown in left of Fig. 3(a) (the solid outer rectangle), we segment the region with the highest disparity as the region of the face. The threshold of disparity of face region is automatically decided, so no need to restrict the distance between face and camera. This method is superior to other methods because it is robust when the background is dynamic or its color is similar to the skin.

Coarse to fine correspondence matching algorithm on the pyramid stereo images is used to speed up matching procedure. First the stereo pictures are applied with pyramidial decomposition, which results in serials of picture with different resolution. The match result of lower resolution is used to instruct match in higher resolution that guarantees a fast speed. Given the disparity histogram of the lowest resolution picture with the peak of highest disparity at $d^*$, and $d(x,y)$ is the disparity at $(x,y)$. We have:

$$
\begin{align*}
    d'(x,y) &= \begin{cases} 
        d(x,y) & d(x,y) \in [d^*-\delta, d^*+\delta] \\
        d^* & d(x,y) \notin [d^*-\delta, d^*+\delta]
    \end{cases} \\
    \text{Mark “right” to this frame and go on tracking} & \quad \text{Mark “wrong” and decide next state “detection”?}
\end{align*}
$$

$$
\begin{align*}
    \text{Are features correct?}
\end{align*}
$$

$$
\begin{align*}
    \text{Detect features by the instruction of global face detection}
\end{align*}
$$

$$
\begin{align*}
    \text{Detect face area in the segmented face region using multiple templates}
\end{align*}
$$

$$
\begin{align*}
    \text{Compute disparity map and segment the foreground as head (face) region}
\end{align*}
$$

$$
\begin{align*}
    \text{“detection”? “tracking”?}
\end{align*}
$$

$$
\begin{align*}
    \text{Stereo cameras}
\end{align*}
$$

$$
\begin{align*}
    \text{Init state to “pure detection”}
\end{align*}
$$
\( d'(x,y) \) is used to supervise the match in next resolution, thus only pixels with disparities in the neighborhood of \( d^* \) in the lowest resolution picture could be better matched in higher resolution pictures. Finally these matched pixels in the highest resolution image are grouped together to be the face region.

In this way face area can be robustly focused from complex dynamic changing input images.

3.2. GLOBAL FACE LOCATING

To accurately locate the face on the result of stereo vision, a simplified face detection algorithm [8] that uses multiple template matching is used. The method in short is as follow:

1. Correlation between Template and Sample Window

Given a template \( T[M][N] \) with the intensity average \( \mu_T \) and the squared difference \( \sigma_T \), an image window \( R[M][N] \) with \( \mu_R \) and \( \sigma_R \), the correlation coefficient \( r(T, R) \) is as follows:

\[
    r(T, R) = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (T[i][j] - \mu_T)(R[i][j] - \mu_R)}{M \cdot N \cdot \sigma_T \cdot \sigma_R}
\]

2. Face Templates.

An average face (Fig.3) is generated from a set of 50 mug-shots, of which each face is rectified by its feature points including pupils and corners of the mouth in geometry, and then is transformed to a normalized scale (50x50) and a gray distribution of the same average (128) and squared difference (64). The average face is subsampled to the size of 20x20 as the fundamental face template and of which an eyes-in-whole template is cut out with the size of 20x8.


At each point of the foreground image, firstly, the upper part of each scanned window is matched by the eyes-in-whole template, and then if the match is over an eyes-in-whole threshold, the whole window is matched by the face template, keeping the maximum one over a face threshold as the final location result [8]. Faces located in a video sequence are given in Fig.3(a) (inner dashed rectangle).

3.3. FACIAL FEATURE DETECTION

Under the instruction of global face size and position, it goes on to local feature detection. The facial feature detected includes: eye centers, nose, and mouth corners.
The global face indicates a rough pattern of organ distribution. In the area found in Sec3.2, where gray-level values change quickly are organ candidate positions. We can also know the rough size of face in that frame according to the global detection result. The gray-level value change at the positions of organs is closely related to face size. The wavelet analysis [9] can detect gray-level change in different scales; Gabor Wavelet Transform [10][11] can detect gray-level changes in different scale and direction. These multi-scale analysis methods compute gray-level changes at all scale and direction, which are computation intensive not suitable for real-time detection and tracking. In fact, what needed in this problem is just the gray-level changes at a specific scale which best represents the organ (“organ-scale”).

We propose a “multi-scale like” analysis with simple “Sobel Filter”. Sobel filter is a fixed filter, which can only detect gray-level changes in some “specific scale range”. Horizontal Sobel works well in segmenting facial organs when face is scaled to a range \([R_{min}, R_{max}]\). This “specific scale range” \([R_{min}, R_{max}]\) is determined by experiment on face at scale from 30pixel to 200pixel and the result is \(R_{min} \approx 80\), \(R_{max} \approx 125\). Given the global face detection result, according to the face area and size it is scaled to \((R_{min}+R_{max})/2\) and convoluted by Sobel filter, as shown in Fig6(a)-(b).

Facial feature detection is at the order “eyes→nose→mouth” as follows.

1. Detect eyes

As shown in Fig6(b), organs have higher gray-level value after Sobel convolution. Suppose the area of organ is \(\alpha\%\) of the whole area of face. The pixels with largest \(\alpha\%\) (See Fig6(c): accumulate histogram map) gray-level values are segmented as organ candidates, as shown in Fig6(d). In our system, \(\alpha = 10\).

Connected components in Fig6(d) without suitable area, shape or position are eliminated from candidates of eyes. The candidates of eyes are shown in Fig7-left. Geometric possible two eyes are matched into a pair of eyes (Fig7-middle). Then, Principle Component Analysis (PCA) based eigen-eyes method [6] is used to find real pair of eyes from candidate pairs of eyes (Fig7-right).

The PCA eigen-eye method is described as follows:

PCA is applied to normalized pair-eye samples (Fig8(a)). Eigen vectors with eigen values sum up to 95% of all eigen values are chosen as eigen eyes (Fig8(b)). Candidate pairs of eyes are projected to eigen eyes and only real pair of eyes can achieve highest reconstruct correlation coefficient (Fig8(c)). Thus, the real pair of eyes is chosen.
2. Detect nose and mouth corners

In Fig.6(d), we have the position of eyes. Nose position is detected referring to eyes position.

Referring the size and position of eyes, the approximate nose and mouth position is shown as the dashed rectangle in Fig.9. The gray-level values are accumulated along “X-axis” and denoted \( X_{\text{nose}} \) at the first peak. Accumulating gray-level value in \([X_{\text{nose}}-\sigma, X_{\text{nose}}+\sigma]\) along “Y-axis” finds the left and right range of nose.

Mouth corners can easily be found in Fig.6(d) according to the position of nose.

4. FACIAL FEATURE TRACKING

Facial feature tracking is based on the features detected correctly in the first frame. Traditional feature tracking need to mark features in the first frame, and then focus research on the tracking strategy [11][12]. In our algorithm, feature in the first frame is detected and verified automatically.

4.1. VERIFYING FACIAL FEATURE DETECTION RESULT

The detection correctness is verified between consecutive frames. Both geometric restriction and rigid body restriction are used in verifying.

1. Geometric restriction

Suppose \( \{ f_1, f_2, f_3, \ldots, f_n \} \) are consecutive \( n \) frames, \( \{ d_{11}, d_{21}, d_{31}, d_{41}, d_{51} \} \) are distance between features \( \text{“left eye-right eye”, “left eye-left mouth corner”, “left eye-right mouth corner”, “left mouth corner-right mouth corner”, “left mouth corner-nose center”} \) detected in \( f_i \)-th frame. It is considered satisfying geometric restriction, if following inequality is satisfied:

\[
\forall i, k \in [1, n], i \neq k, \quad |d_{i,j} - d_{k,j}| < \delta, \quad i = 1, 2, 3, 4, 5
\]

2. Rigid body restriction

Face motion with little expression can be considered as rigid body motion. The distance between camera and face is much larger than depths between facial organs, so we can regard all facial features in a plane. The motion is simply represented by affine transform. Suppose \( \textbf{X} \) is original features and \( \textbf{X}' \) are features after motion, it has:

\[
\textbf{X}' = \textbf{RX} + \textbf{T}, \quad \textbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}, \quad \textbf{T} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}
\]

3 points pair \( (\textbf{X}, \textbf{X}') \) can completely decide the transform parameter \( \textbf{R} \) and \( \textbf{T} \). With detected 7 points, we have \( C_7^3 = 35 \) possibilities to select 3 points. The transform parameters between 35 groups of 3 points are computed: \( \textbf{R} : \{ \textbf{R}_1, \textbf{R}_2, \ldots, \textbf{R}_{35} \} \) and \( \textbf{T} : \{ \textbf{T}_1, \textbf{T}_2, \ldots, \textbf{T}_{35} \} \), and all detected points are computed to test the rigid body restriction as follows:

\[
\max ||\textbf{X}' - (\textbf{R}\textbf{X} + \textbf{T})|| < \delta, \quad \text{where} \quad \delta \quad \text{is error control parameter.}
\]

Consecutive \( n \) frames \( \{ f_1, f_2, f_3, \ldots, f_n \} \) which all satisfy geometric and rigid restriction are mark with “feature correct”, and detection result of the \( f_n \)-th frame is set as the reference, which is used to supervise further tracking.

4.2. TRACKING STRATEGY
At the pre-condition that features in the i-th frame are correct, features in the (i+1)-th frame are tracked based on last frame. The details are as follows:

1. Referring to the features of the i-th frame, the approximate position in the (i+1)-th frame is decided
2. Method described in Sec3.2 is used to detect features.
3. Verification feature using both geometric restriction and rigid body restriction. If it is right, this frame is marked “right”, and reference is updated; or, marked “wrong”.

If several consecutive frames are marked “wrong”, system goes back to pure detection.

5. RESULTS OF DETECTION AND TRACKING

We test detection and tracking performance on 189 video sequences that each with 50 frames. Clients focus their view around a circle towards the cameras for including various poses. In all the frames in 189 video sequences, the number of frames marked correct is 9280 frames. The correct rate of facial feature detection is 9280/(189*50)=98.2%. It achieves 25fps detection and tracking speed on a PIII966MHz PC.

Fig. 10 shows detection and tracking results in two video sequences.

Fig. 10(a) Video 1: the 1st, 5th, 10th, 15th, 19th, 24th, 29th, 33rd, 37th, 41st, 45th, 50th frame
6. CONCLUSION

In this paper, a “coarse-to-fine” facial feature detection and tracking system under complex background is presented. Stereo cameras are used to roughly and quickly segment face area from complex background. Then, the multiple template matching method is used for fine face location. Facial organ candidates are extracted from the detected face region at a specific scale space called organ scale for Sobel filter. Finally, eyes, nose and mouth corners are detected. Techniques for checking and correcting errors in facial feature detection based on multiple cues are developed to make the algorithm more robust in facial feature detection and tracking in video sequence.

REFERENCES