Tracking and Recognition of Multiple Faces at Distances

Rong Liu, Xiufeng Gao, Rufeng Chu, Xiangxin Zhu, and Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun Donglu Beijing 100080, China

Abstract. Many applications require tracking and recognition of multiple faces at distances, such as in video surveillance. Such a task, dealing with non-cooperative objects is more challenging than handling a single face and than tackling a cooperative user. The difficulties include mutual occlusions of multiple faces and arbitrary head poses. In this paper, we present a method for solving the problems and a real-time system implementation. An appearance model updating mechanism is developed via Gaussian Mixture Models to deal with tracking under head rotation and mutual occlusion. Face recognition based on video sequence is then performed to get the identity information. Through fusing the tracking and recognition information, the performance of them are both improved. A real-time system for multi-face tracking and recognition at distances is presented. The system can track multiple faces under head rotations, and deal with total occlusion effectively regardless of the motion trajectory. It is also able to recognize multi-persons simultaneously. Experimental results demonstrate promising performance of the system.

1 Introduction

Many applications need the system to have the ability to track and recognize faces at a distance, for example, in intelligent video surveillance. In such settings, the system should be able to keep track of the faces when the people are not facing to the camera. In addition, mutual occlusions may occur as multiple faces move and interact one another, and then some faces may disappear for several frames due to total occlusion. Moreover, the head poses of the persons in the scene are very free. As a result, tracking and recognition of multiple faces at distances is a challenging task.

While there is an abundant literature on face tracking in video sequences, not much is focus on developing a system for the above mentioned problem. To handle the head rotation, Chen and Kee [3] train a head detector based on head shapes. However, the high false ratio of this head detector leads it can't be used directly as a face detector in video surveillance. Yang and Li [5] present a system which can track varying poses. Due to them only use the face detection results to update the tracker, it is not very stable when the person turns back from the camera. Multi-features are also fused together to keep more stable face tracking [4, 7], such as contour, color and motion information. To develop a real-time system, however, the computation complexity is generally high.

To handle the total occlusion, Niu et al [6] use Kalman filter to predict the motion trajectory. The defect is it can not endure too free motions. Lerdsudwichai et al [8] use

color histogram to build the upper body model to recover the person after occlusion. It is easy to be failure, when the occluded person reappears with her/his face turning away from the camera.

Based on face tracking, the sequence information of each face can be easily obtained. Taking advantage of these video sequences, face recognition can be more accurate through Video-based recognition method. There are some works related to it. Zhao and Chellappa et al [1] analyze the advantages of face recognition based on video. McKenna and Gong et al [9] model face eigenspace in video data via principal component analysis. Probability vote approach is then used to fuse the sequence information. Krueger and Zhou et al [10] take advantage of the temporal information to improve the recognition performance. These recognition methods are initially developed to recognize one person in video sequence. In addition, how to fuse the temporal and identity information for recognizing multi-faces is still a problem. In [2], two cameras, a static and a PTZ, work cooperatively. The static camera is used to take image sequences for face tracking, and the PTZ camera is used to take images for face recognition. In this way, the system is supplied with high quality images for face recognition since the PTZ camera can be adjusted to focus on the face to be recognized. However, that system can only recognize a single face in the scene.

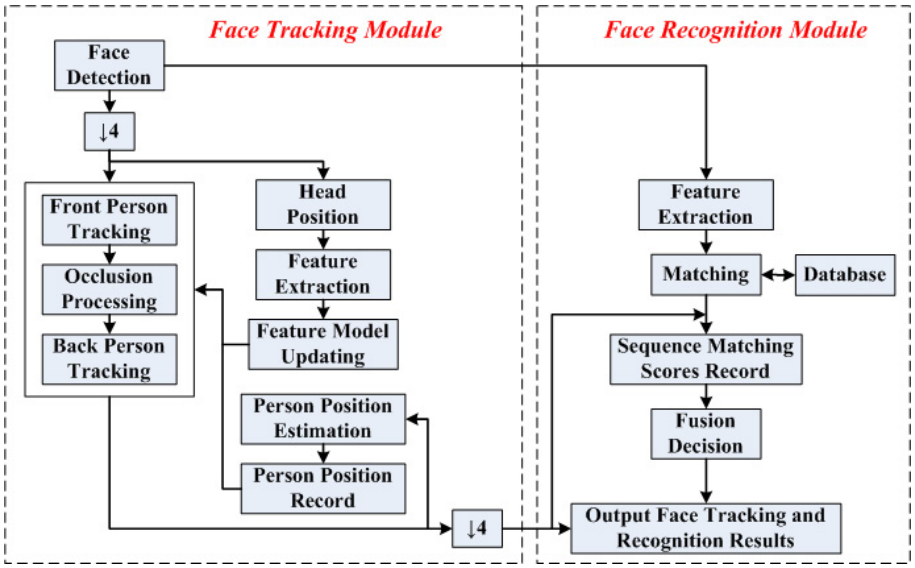
In this paper, we develop a method and a system for tracking multiple faces at distances (see Fig. 1). The method involves a face tacking module and a face recognition module. In the face tracking module, Two GMMs are used to represent the appearance of each person. One is applied to the head appearance to keep head tracking. The other is applied to that of the upper body to deal with occlusions. These two models are updated online.

In the face recognition module, a face recognition engine, which is learned from a classifier based on a Local Binary Pattern (LBP) representation [13] using AdaBoost learning [14], is used to obtain identity matching scores for each frame. These matching scores are computed over time to obtain a score sequence. The matching scores are fused and used to associate the tracked persons in consecutive frames, as well as to provide face recognition results. When the fused scores are below a given threshold, the system will consider the corresponding persons as unenrolled ones.

Based on the two modules introduced above, a real-time system for multiple faces tracking and recognition at distances is built. The system can track multiple faces under head rotations, and deal with total occlusion effectively regardless of the motion trajectory. It is also able to recognize multi-persons simultaneously.

The remainder of this paper is organized as follows: Section 2 develops an updating mechanism for appearance model based on Gaussian Mixture Models to deal with tracking under head rotation and mutual occlusion. Section 3 describes the recognition method used in the system. Section 4 describes how to fuse the tracking and recognition information to improve the performance of the system. Section 5 presents experimental results.

(a)



(b)

Fig. 1. The system: (a) environment setting, and (b) diagram.

2 Face Tracking Method

In this section, we describe the tracking module for handling head rotation and total occlusion. Feature modeling and online model updating are two important components in object tracking. Two GMMs are learned to model the colors of the head and upper body of each tracked people, one for tackling head rotation and the other for handling total occlusion. An online GMM updating mechanism is incorporated.

2.1 GMM and Online Updating

A GMM with K component densities at time t can be modeled as follow:

$$P(X_t \,|Y\,) = \sum_{k=1}^{K} w_{k,t} \cdot N(X_t, \mu_{k,t}, \Sigma_{k,t}) \tag{1}$$

where X_t denotes the appearances of the person in the video sequence and Y denotes the tracked person. $N(X_t, \mu_{k,t}, \Sigma_{k,t})$ denotes the k-th Gaussian component with mean vector $\mu_{k,t}$ and covariance matrix $\Sigma_{k,t}$. $w_{k,t}$ is the corresponding weight. The appearance information is the head colors when the GMM model is applied to head tracking. It describes the colors of the upper body when the GMM is used to deal with occlusion.

During track process, new color information of tracked person can be obtained in each frame. Let x_{t+1} be the color information of tracked person Y obtained at time $t+1$. It can be modeled as

$$P(x_{t+1}\,|Y\,) = \sum_{k=1}^{K'} w'_{k,t+1} \cdot N(x_{t+1}, \mu'_{k,t+1}, \Sigma'_{k,t+1}) \tag{2}$$

Due to the occlusion or the interference of background, more color components are needed to describe the model $p(x_{t+1}\,|Y)$. So the K' in Equ. 2 is always larger than the K in Equ. 1.

$p(x_{t+1}\,|Y)$ is used to update $p(X_t\,|Y)$ into $p(X_{t+1}\,|Y)$. However, some components of the distribution should not be used for the update, such as those belonging to occluding objects and the background. The distribution distances between these and each component of old model $p(X_t\,|Y)$ are usually great. The components of $p(x_{t+1}\,|Y)$, which have big Mahalanobis distances to $p(X_t\,|Y)$, are dropped in the updating process. Only those components of $p(x_{t+1}\,|Y)$, which have small Mahalanobis distances to $p(X_t\,|Y)$, are used to update $p(X_t\,|Y)$. The updating formulae are as follows:

$$w_{k,t+1} = \frac{w_{k,t} + \sum_{i=1}^{K'} I_{i,k} w'_{i,t+1}}{1 + \sum_{i=1}^{K'} I_{i,k} w'_{i,t+1}} \tag{3}$$

$$\mu_{k,t+1} = \frac{w_{k,t}\mu_{k,t} + \sum_{i=1}^{K'} I_{i,k} w'_{i,t+1}\mu'_{i,t+1}}{w_{k,t} + \sum_{i=1}^{K'} I_{i,k} w'_{i,t+1}} \tag{4}$$

$$\Sigma_{k,t+1} = \frac{w_{k,t}\left[\Sigma_{k,t} + (\mu_{k,t+1} - \mu_{k,t})(\mu_{k,t+1} - \mu_{k,t})^T\right] + \sum_{i=1}^{K'} I_{i,k} w'_{i,t+1} B_i^{(k)}}{w_{k,t} + \sum_{i=1}^{K'} I_{i,k} \cdot w'_{i,t+1}} \tag{5}$$

where $B_i = \left[\Sigma'_{i,t+1} + (\mu_{k,t+1} - \mu'_{i,t+1})(\mu_{k,t+1} - \mu'_{i,t+1})^T\right]$, $I_{i,k}$ is an indicator. $I_{i,k} = 1$, if the i-th component of $P(x_{t+1}\,|Y\,)$ is matched to the k-th component of $P(X_t\,|Y\,)$. Otherwise, it is equal to zero.

The process of head rotation can be considered as a process of color distribution change in tracking. In this process, the online update method presented above plays two roles: (1) Accepts the matched color model to update the former one. It can maintain the adaptability of the feature model. (2) Prevents the unmatched color model from participating in the updating process. It keeps the validity of the feature model.

After the feature model is obtained at current time, we can construct a weight map for mean shift tracking based on the GMM. To reduce effect of background, the ratio of the head appearance model to the one of background is used to construct the weight map [5]. This treatment is proved effective.

2.2 Handle the Total Occlusion

Fig. 2 shows the area used for handling occlusion. This area is modeled to compare with the upper body models which is built by GMM and updated by the method presented in section 2.1. The person whose upper body model is less matched to the occlusion area is considered as the back person. The Chi-square (χ^2) statistic is used to measure the distinction between the models.

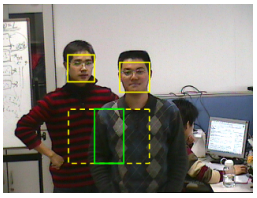


Fig. 2. Each upper body area, which is shown in dashed rectangle, is obtained relative to the face position. The intersection area of the upper body areas is considered as the area of occlusion (green rectangle).

When occlusion occurs, the information of the front person keeps intact. Therefore the position of her/him can be considered as prior knowledge using the general tracking method. Under the tracking of the front person, the weight of the area which is consistent to the moving direction is enhanced. After obtaining the position of the front face, the weight map of the back person is reconstructed. In this map, the weight in the head area of the front person is reduced. As a result, the tracking of the back person will not be affected by the front person (see Fig. 3).

In Fig. 3, the position relationship is estimated in frame 159, and then the weight adjustment for the back person is done in the next frames until the occlusion ends.

3 Face Recognition Method

The face recognition engine is learned from a classifier based on a Local Binary Pattern (LBP) representation [13] using AdaBoost learning [14].

Recently, Local Binary Patterns(LBP) is introduced as a powerful local descriptor for microfeatures of images [13]. The LBP operator labels the pixels of an image by thresholding the 3×3-neighborhood of each pixel with the center value and considering

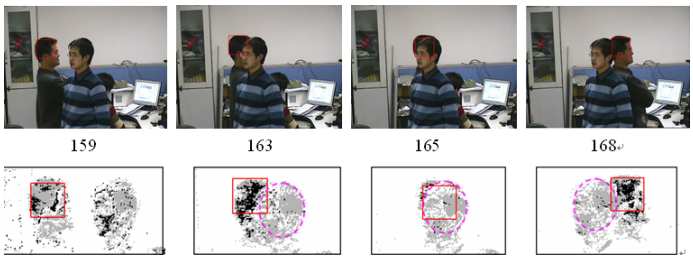


Fig. 3. Face tracking under total occlusion: frames in video sequence (upper), and weight maps for tracking (lower). The red rectangle shows the head position of the back person obtained by tracking mechnism. In the weight map, weight in the ellipse area is reduced. The weights are divided into three levels: high weights (black), middle weights (gray) and low weights (white)

the result as a binary number (or called LBP codes). In our work, a histogram of the base LBP codes is computed over a local region centered at each pixel, and it is considered as a set of individual features. Given a training set of LBP histogram features of faces subjected to image noises, slight pose changes and alignment errors, the AdaBoost learning method can find the most discriminative features and thereby build a strong classifier. Due to AdaBoost is usually used to solve two-class problem, we convert the multi-class problem to many two-class problems using intra-person and extra-person divergence [15].

The recognition engine described above performs one-to-many matching and outputs a matching score for each person in a database. The person's identity is determined based on a fused matching score, which is the average of several consecutive matching scores. The fused score is compared with a threshold, then a decision is made to obtain the person's identity referenced in the face database, or label him/her as unenrolled.

4 Fusing the Tracking and Recognition Information

Two types of IDs are maintained in a track list for tracking and recognition of each face: the TID (Tracking ID) and the RID (Recognition ID). The TID associates a face in different frames, whereas the RID identifies a face referenced in the database. The TID is obtained from the tracking module, whereas the RID is obtained as the result of the aforementioned face recognition decision.

After the TID and the RID are both obtained, they are bonded together, and putted into a track list. For face tracking, the use of RIDs makes face association in different frames as 1-1 matching problem. For recognition, TIDs are used as index for face identity. The tracked person is deleted from the list after the person leaves the scene.

The process of fusing tracking and recognition information is shown in Fig. 4. Using the track list, we can obtain a TID-indexed face and a RID-indexed face for each tracked person. Then a 1-1 matching is made between both of them. If they are matched,

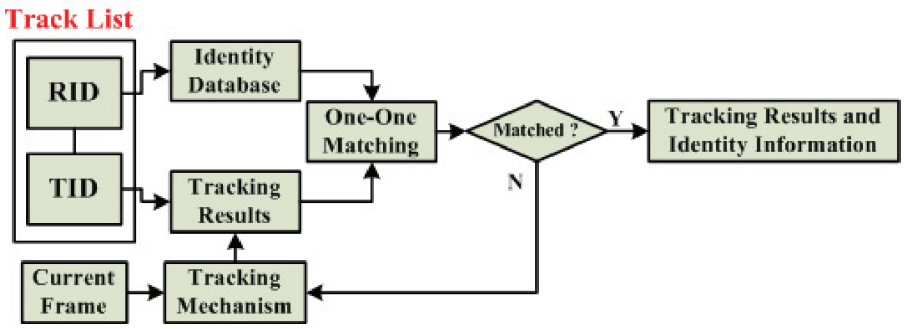


Fig. 4. The diagram of using TID and RID to fuse tracking and recognition information

we accept the tracking result and the corresponding identity information. Otherwise, we reduce the weight of this face area, and track it again. Through this process, the performance of tracking and recognition are both improved.

5 Experimental Results

The system is implemented on a standard PC (Pentium IV at 3.0GHz). The video image size is 640x480 (24 bits per pixel) captured by IKEGAWA SN-600H-22 at 25fps and it is down-sampled to 320x240 for face tracking (while face recognition is performed on the resolution of 640x480). The RGB color space is down-sampled to 20x20x20 bins for building the color distribution of the object. The face is cropped to the size of 120. The tracking mechanism is initialized by multi-view face detector [11]. An ellipse model matching method [4] is used to find the head region from the position of detected face. The system works at about 10 fps. The following presents experiments for face tracking and face recognition.

The system was tested with significant head rotations in-plane and out of plane, large scale changes, multiple persons, nonlinear fast moving and total occlusion. We deliberately selected clips taken under difficult conditions, especially those with rotation and occlusion. Fig. 5- 7 present the results.

In Fig. 5, the person looked up at frame 63; looked down at frame 69; turned about at frame 132; turned back at frame 139. These were the most challenge poses when tracking under head rotation. The high performance of the system to handle head rotations is shown in this experiment.

In Fig. 6, the person was occluded by hand from frame 87 to frame 89, and by arm from frame 119 to frame 121. For a tracking method based on color feature, occluded by the object with the similar color is a great challenge. In this experiment, the hand which is of the skin color was used to occlude the face. It shows that the system can robustly handle part occlusions.

In Fig. 7, the person with white cloth was first occluded by the person with red cloth from frame 176 to frame 180. When the total occlusion ends, the back person also kept profile pose which can not be detected by face detector. From frame 187 to frame 189,

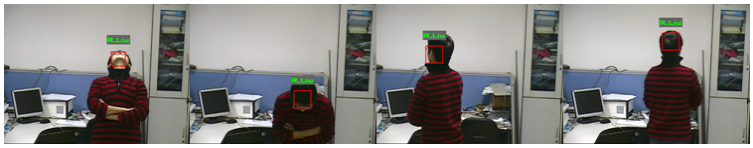


Fig. 5. Several typical head poses: looking up (frame 63), looking down (frame 69), turning about (frame 132) and turning back (frame 139). The red rectangle shows the position of head obtained by tracking mechanism. The person's identity is obtained by recognition, and showed above face.

Fig. 6. Part occlusions: occluded by hand (from frame 87 to frame 89), and by arm (from frame 119 to frame 121). The red rectangle shows the position of face obtained by tracking mechanism. The person's identity is obtained by recognition, and showed above face.

the person with white cloth was occluded for a long time, and moved to the inverse direction after occlusion ends. These instances are hard to be handled in a single face tracking system. It shows that the system can robustly handle total occlusions.

The recognition performance of the system was tested in a indoor environment. In the face recognition process, the system were done in the form of one-to-many matching in each frame and fusion decision based on the matching scores of past frames, with the following protocol: fifty people were enrolled as clients, with twenty templates per person recorded. Images of the fifty clients were not included in the training set. The enrolled population was mostly composed of Chinese people with a few Caucasians and Negroes. Five people participated as the regular imposters (not enrolled) and some visitors were requested to participate as irregular imposters.

These participants entered the scene of the system several times for test. It provided statistics of correct recognition rate, correct rejection rate and average recognition time. Some client people deliberately challenged the system by exaggerated expressions, turning back to the camera or occluding part of the face with a hand so that the system did not recognize them. We counted these as invalid. Only those tests which were reported having problems getting recognized were counted as false rejections. On the other hand, the imposters were encouraged to challenge the system to get false acceptances.

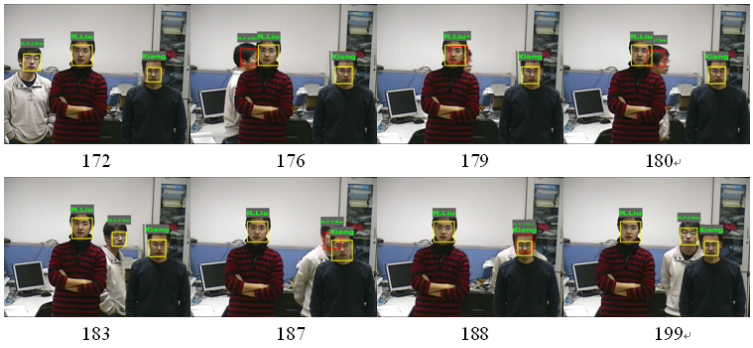


Fig. 7. Multi-person demos. There are three persons in the scene. The person with white cloth is first occluded from frame 176 to frame 180, and then from frame 187 to frame 199. The rectangles show the positions of faces. The identity of each person is obtained by recognition, and showed above each face.

After near seven hundred tests, the system has demonstrated excellent accuracy, speed, usability, and stability. The correct recognition rate is above 99%. The correct rejection rate is above 97%. the ART (Average Recognition Time) is below 2 seconds. Hence, we can conclude that the system achieves high-performance of recognition.

6 Conclusion

This paper presented a method for tracking and recognition multiple faces at distances, and its real-time system implementation. The method employs several tracking strategies for reliable face tracking and further enhance it by making use of face identity information. The system can track multiple faces under head rotations, and deal with total occlusion effectively regardless of the motion trajectory. It is also able to recognize multiple persons simultaneously and the recognition performance is demonstrated accuracy, speed, usability, and stability.In the future, we will extend the work to tackle a more challenging problem, that of more reliable face recognition at distances, by taking advantages of the present tracking system and developing a face recognition module optimized for the surveillance ID task.

Acknowledgements

This work was supported by the following funding resources: National Natural Science Foundation Project #60518002, National Science and Technology Supporting Platform Project #2006BAK08B06, National 863 Program Projects #2006AA01Z192 and #2006AA01Z193, Chinese Academy of Sciences 100 people project, and the Authen-Metric Collaboration Foundation.

References

1. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: a literature survey. ACM Computing Surveys. **35** (2003) 399-458
2. Prince, S., Elder, J., Hou, Y., Sizinstev, M., Olevskiy, E.: Towards face recognition at a distance. In Proc. IET Conference on Security (2006)
3. Chen, M.L., Kee, S.: Head tracking with shape modeling and detection. Computer and Robot Vision (2005) 483-488
4. Birchfield, S.: Elliptical head tracking using intensity gradients and color histograms. Computer Vision and Pattern Recognition (1998) 232-237
5. Yang, T., Li, S.Z., Pan, Q., Li, J., Zhao, C.: Reliable and fast tracking of faces under varying pose. Automatic Face and Gesture Recognition (2006) 421-426
6. Niu, W., Jiao, L., Han, D., Wang, Y.F.: Real-time multiperson tracking in video surveillance. AInformation, Communications and Signal Processing (2003) 1144-1148
7. Jin, Y.G., Mokhtarian, F.: Towards robust head tracking by particles. Image Processing (2005)
8. Lerdsudwichai, C., Abdel-Mottaleb, M., Ansari, A.: Tracking multiple people with recovery from partial and total occlusion. Pattern Recognition. **38** (2005) 1059-1070
9. McKenna, S.J., Gong, S., Raja, Y.: Face recognition in dynamic scenes. British Machine Vision (1997)
10. Krueger, V., Zhou, S.: Examplar-based face recogntion from video. Automatic Face and Gesture Recognition (2002)
11. Li, S.Z., V., Zhu, L., Zhang, Z.Q., Blake, A., Zhang, H.J., Shum, H.: Statistical learning of multi-view face detection. European Conference on Computer Vision, Copenhagen, Denmark (2002)
12. Li, S.Z., R.F. Chu, S.C. Liao, L. Zhang.: Illumination Invariant Face Recognition Using Near-infrared Images. IEEE transaction on Pattern Analysis and Machine Intelligence. to appear April 2007.
13. Ahonen, T., Hadid, A., Pietikainen, M.: Face Description with Local Binary Patterns: Application to Face. Recognition. IEEE transaction on Pattern Analysis and Machine Intelligence. **28** (2006) 2037-2041.
14. Viola, P., Jones, M.: Robust Real-time Object Detection. International Journal of Computer Vision. **57** (2004) 137-154.
15. VioMoghaddam, B., Nastar, C., Pentland, A.: A Bayesian similarity measure for direct image matching. Media Lab Tech Report No.393. **57** MIT(1996)