Cap Detection for Moving People in Entrance Surveillance

Rui Min
Institut EURECOM
Sophia Antipolis, France
(+33) 04 93 00 82 69
min@eurecom.fr

Jean-Luc Dugelay
Institut EURECOM
Sophia Antipolis, France
(+33) 04 93 00 81 41
jld@eurecom.fr

ABSTRACT
While there has been an enormous amount of research on face recognition under pose/illumination changes and image degradations, problems caused by occlusions are mostly overlooked. Moreover, most of the existing approaches of face recognition under occlusion conditions focus on overcoming facial occlusion problems due to sunglasses and scarf. To the best of our knowledge, occlusion due to cap has never been studied in the literature, but the importance of this problem should be emphasized since it is known that bank robbers and football hooligans take advantage of it for hiding their faces. This paper presents a solution to this newly identified face occlusion problem – the time-variant occlusion due to cap in entrance surveillance, in the context of face biometrics in video surveillance. The proposed approach consists of two parts: detection and tracking of occluded faces in complex surveillance videos; detecting the presence of cap by exploiting temporal information. The detection and tracking part is based upon body silhouette and elliptical head tracker. The classification of cap/non-cap faces utilizes dynamic time warping (DTW) and agglomerative hierarchical clustering. The proposed algorithm is evaluated on several surveillance videos and yields good detection rates.

Categories and Subject Descriptors
I.4.8 [**Scene Analysis**]: Object recognition, Tracking, Time varying imagery; I.2.10 [**Vision and Scene Understanding**]: Video Analysis; I.4.9 [**Applications**].

General Terms
Experimentation, Algorithms, Security

Keywords
Entrance surveillance, face recognition, occlusion detection, security management.

1. INTRODUCTION
Face recognition, the least intrusive biometric technique from the sampling point of view, has been applied to a wide range of commercial and law enforcement applications. With the emphasis on real world scenarios (e.g. face recognition in video surveillance), in the last decade, an enormous amount of research on face recognition under pose/illumination changes and image degradations has been conducted. However, problems caused by occlusions are mostly overlooked, although facial occlusions are quite common in non-cooperative systems such as video surveillance applications.

Facial occlusions, which might also occur for benign behaviors, are often related to several severe security issues. For example, ATM criminals and football hooligans tend to wear scarf, sunglasses and/or cap to prevent their faces been recognized. Therefore, face recognition robust to partial occlusions is an essential technique for security management and forensics. Among most of the existing solutions, local features and local components based methods (e.g. [1][2]) become prominent, since locality emphasized algorithms are less sensitive to partial occlusions in comparison with conventional holistic algorithms. Nevertheless, information from occluded parts can still hinder the recognition performance in those approaches. More recently, researchers have revealed that prior knowledge of occlusions can significantly improve the accuracy of local feature based face recognition [3][4], since the harmful information from occluded parts can thus be excluded explicitly. Hence, accurate occlusion detection is crucial to achieve occlusion robust face recognition.

It should be noted that the state-of-the-art of occluded face recognition addresses the time-invariant occlusions only (e.g. scarf, sunglasses or artificially generated occlusions) from a single image. Time-invariant indicates that the occlusion is not changing without external forces within a certain period of time. In this paper, we identify a new type of occlusion: the time-variant occlusion, specifically speaking the occlusion due to cap in entrance surveillance (though headdresses like hard hat in [5] is not considered as an occlusion problem). Nowadays, Closed-Circuit Television (CCTV) cameras are deployed everywhere for security purpose. The most common deployment of such cameras

Figure 1. Illustration of Entrance surveillance scenarios

is at the room ceiling in order to monitor the entrances of various places (e.g. banks, ticket machine of subway station, supermarkets, libraries, and stadium) (see Figure. 1). With videos captured by such cameras, the identities of people inside the restricted place can be recorded and later recognized by automatic face recognition (AFR) systems or human observers. However,

*Area Chair: Tian-Tsong Ng

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

according to many recent police reports, bank robbers, shop thieves and football hooligans usually wear a cap when entering the places where they commit crimes. Due to the viewing angle problem, a close-distance face captured by a CCTV camera is often occluded by the visor, whereas a far-distance face is not occluded but it has a low resolution. The occlusion caused by cap with a visor is therefore varying along with people moving. Because of the trade-off between occlusion and image quality, obtained face images are usually unrecognizable from the recorded videos. Thus, it is imperative to equip automatic cap detection systems in entrance surveillance, which cannot only detect suspicious persons, but also provide prior knowledge of occlusion to improve the face recognition of criminals.

In this paper, we address then the occlusion detection due to cap for entrance surveillance. The proposed approach consists of first detect and track the face region and then detect the presence of cap within the tracked face. The detection and tracking part is based upon silhouette analysis and elliptical head tracker [6]. The classification of cap/non-cap faces adopts dynamic time warping (DTW) and agglomerative hierarchical clustering.

2. CHALLENGES AND INNOVATIONS

There are many challenges for the proposed cap detection system. First of all, in entrance surveillance, even a close-distance face is relatively far away from the camera (in comparison with the traditional biometric scenario). In addition, various variations due to occlusion, rotation, low resolution and complex backgrounds make the correct detection of face region a difficult problem. In the targeted entrance surveillance scenario, the most common approach – Viola-Jones algorithm [7] implemented in OpenCV cannot even find a non-occluded face correctly (mainly due to the resolution problem). For this reason, we implemented a customized method to detect and track the face region, which exploits more robust features such as body silhouette and head geometry.

Secondly, occlusion due to cap is time-variant. Unlike the time-invariant occlusions (such as scarf and sunglasses), the occlusion caused by cap is varying along time. When a face is approaching, the occluded area on the face increases accordingly. The changing of occluded area mainly depends on the speed and pause of people walking. Since the walking habit is non-homogeneous and largely different among individuals, the occlusion situation is also diverse for different people or even the same people in different videos. In addition, the occluded area is also changing because of the rigid head movements (e.g. head nodding and looking way). For above reasons, occlusion detection using a single image is very unlikely to get accurate results. In our approach, we exploit all the frames in the video to reduce the effects of different variations and tracking errors; furthermore, the temporal information is preserved by using DTW to distinguish the characteristics of occlusion varying from the occluded and non-occluded faces.

Last but not least, the proposed system is supposed to work in real time, since the computational capability is limited in distributed sensor networks. The flowchart of the proposed system is given in Figure 2. The simplicity of each of the steps ensures the overall performance of our system.

3. HEAD DETECTION AND TRACKING

The customized head detection and tracking is based on silhouette image and elliptical head tracker. The accuracy of detection and tracking thus largely depends on the quality of extracted body silhouette. Following the well established silhouette analysis in gait recognition [8]; our system is able to achieve good detection and tracking of non-optimal faces in textured background.

3.1 Background Modeling

The first step is to construct a reliable background model for silhouette extraction. Here, the LMedS (Least Median of Squares) method [9] is applied to build the background from a small portion of video sequences, thanks to its power of filter moving objects.

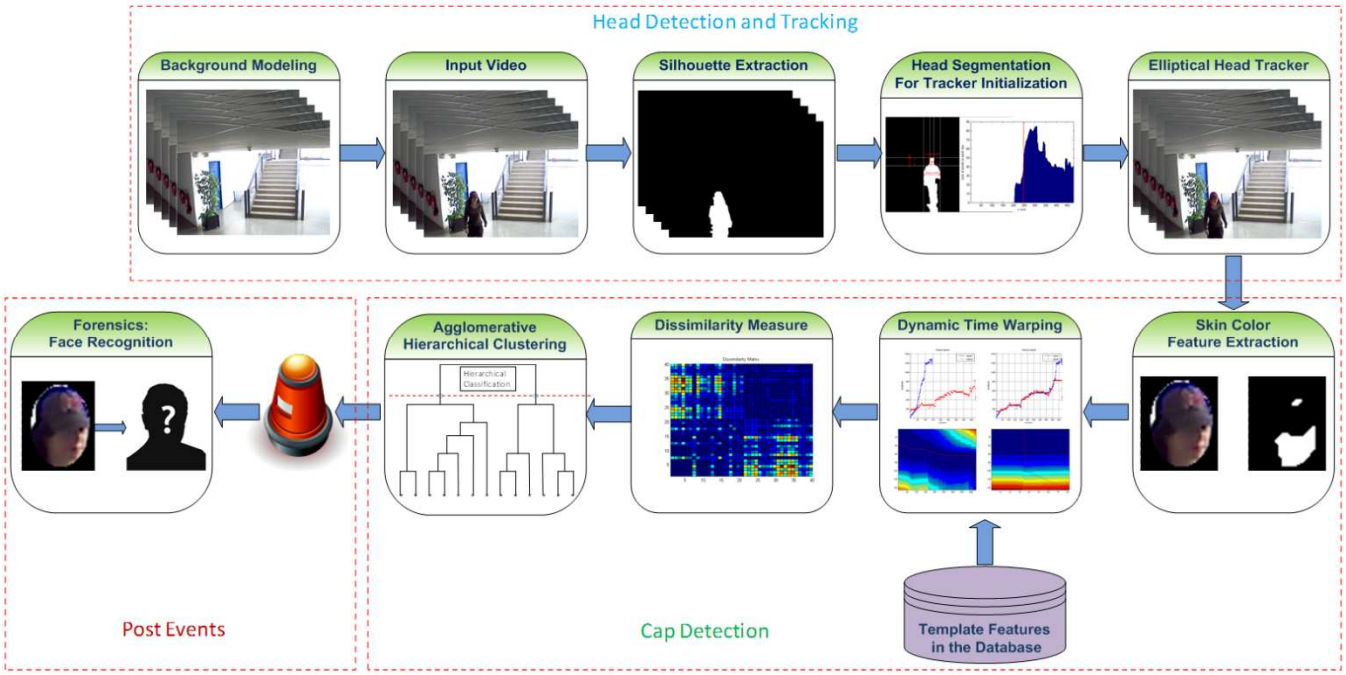

Figure 2. Flowchart of the proposed cap detection system

3.2 Silhouette Extraction

Once the background image is correctly constructed, the foreground detection can be achieved by differencing and thresholding between the background image and the current frame. Because the background subtraction step may introduce noise and distortions in the segmented foreground, post-processing including morphological operations and connected component selection are applied in order to obtain a clear body silhouette.

3.3 Head Segmentation

Correct segmentation of head from the body silhouette is critical to initialize a head tracker. The bounding box of walking body is naturally obtained from the silhouette. To segment the head from the other part of the body, we apply a horizontal projection to the silhouette image. Then the head-shoulder detection is achieved via selecting the histogram steep rise.

3.4 Scalable Elliptical Head Tracker

The tracking part is directly adopted from [6]. The choice of the elliptical head tracker is based on the fact that it is robust to occlusion, rotation, tilting and textured background to some extent. In addition, the merits of scalability, reacquisition and low complexity encourage us to integrate it into our system. However, instead of applying the elliptical head tracker to a gradient map, we fit the ellipse to the body contour which is extracted by a Canny edge detector from the silhouette image. The reason is due to the gradient map of a low resolution head is indistinguishable from the textured background. In the prediction part, the monotonic increment assumption is made. This assumption can accelerate the system speed as well as automatically exclude the person who is leaving the entrance (in which case the face is invisible and also redundant, it was recorded during entering).

In our experiments, the elliptical head tracker correctly tracked the head in 84% of all the frames (of 40 video clips) for 10 different people with and without cap under 2 illumination conditions (morning and evening). Errors are due to incorrect ellipse fitting (23%) or lost tracking (77%).

4. Cap Detection

Because facial components, such as eyes, nose and mouth, are almost undetectable in the entrance surveillance scenario, the only reliable feature for occlusion classification is the skin color. In this paper, a novel Spatial-Temporal (ST) feature representation based on skin color is proposed. Inspired by the work from [10], DTW and agglomerative hierarchical clustering are employed to classify the unsynchronized ST features.

4.1 Skin Color based Feature Extraction

A traditional way [11] to extract skin pixels which is fast and stable is applied. Skin pixels are differentiated from the non-skin pixels in the $YCbCr$ color space by pre-defined thresholds. In this paper, we tested two types of features: 1. the number of skin pixels in the region of interest (ROI); 2. the ratio of number of skin pixels to number of all pixels in the ROI. Then the extracted features from all frames of one video clip are concatenated to construct a ST representation. The perceptual classifications between features extracted from the occluded and non-occluded video clips are shown in Figure 3.

4.2 Dynamic Time Warping

Dynamic time warping, also called curve registration, is widely used in various applications (e.g. speech recognition, musical information retrieval and bioinformatics), thanks to its capability of measuring the similarity between two sequences varying in time or speed. The optimal match between two feature vectors is

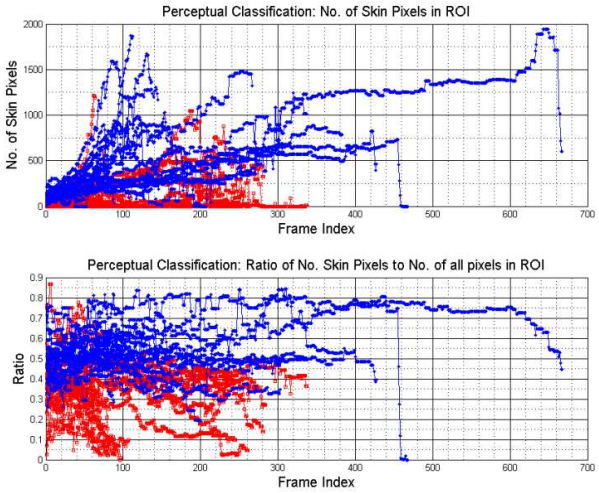

Figure 3. Perceptual classification of the proposed features (Blue: faces without cap, Red: faces with cap)

found by stretching and compression under the constraint of monotonicity. First, a difference matrix $D[i,j] = |Q[i] - T[j]|$ of the query feature vector $Q[i]$ and the template feature vector $T[j]$ is calculated. Then a cost matrix M is recorded by keeping a running tab on the dissimilarities of elements while summing up to a minimum accumulated cost measure via dynamic programming [10]. The formation of M can be written as:

$$M[i,j] = min\begin{pmatrix} M[i-1,j-1] \\ M[i-1,j] \\ M[i,j-1] \end{pmatrix} + D[i,j] \qquad (1)$$

With matrix M, a minimum cost path can be retraced along diagonal. The final dissimilarity between the query and template is thus the summation of all costs on the path, normalized by the length of the longer feature vector between Q and T.

4.3 Dissimilarity Matrices

In our cap detection system, a number of features extracted from video clips of occluded and non-occluded faces serve as the templates. When a query video comes in, the dissimilarities between the query and all templates are computed via DTW. To reduce computation, the dissimilarities between all pairs of templates are pre-calculated and stored as dissimilarity matrices in the database. The system also support online updating by integrating the dissimilarities computed from the query into the matrices. A visual presentation of the dissimilarity matrices is shown in Figure 4.

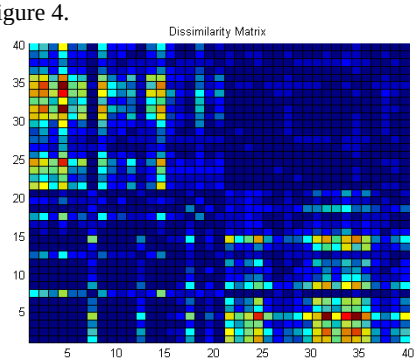

Figure 4. Dissimilarity Matrices of 40 video clips (1-20 without cap, 21-40 with cap, cold color indicates low dissimilarity)

4.4 Agglomerative Hierarchical Clustering

Knowing the dissimilarity matrices, the agglomerative hierarchical clustering algorithm is applied for classification. A dendrogram is returned by the clustering algorithm. In the dendrogram we choose the level with two clusters for decision making. The class label of each cluster is decided by the number of template features within the same cluster; as an example, a cluster belongs to the cap class when there are more template features from cap videos in the cluster and vice versa. A query feature is thus classified according to the cluster's class label. Ward's method is chosen empirically for the clustering task.

4.5 Post Events

Here we briefly discuss the post events after cap detection. In the entrance surveillance scenario, the presence of cap is undesirable for security management. In the cooperative case (e.g. in airport security check), the system can provide information so that guards can kindly ask the passenger to remove his/her cap. In the non-cooperative case (e.g. in bank), it can identify the suspicious person so that increase the security level or draw the attention from a human observer. For forensics purpose from recorded videos, it can be easily combined with local feature based face recognition system as described in [4].

5. EXPERIMENTS

In this paper, we built a rather challenging dataset to simulate the environment of entrance surveillance in real-world. 40 videos (length varying from 60 to 666 frames) have been recorded using an Axis P1343 IP camera fixed at the room ceiling in an indoor terrace (with sunlight during the daytime and artificial lights in the evening). 10 people were asked to walk toward to the camera and enter the door of office below the camera. Each person had twice recording, one time wearing a cap and another time without wearing a cap. The caps have different colors and textures in order to maximize the variations. The experiments are organized in 2 sessions, one session in the morning and another one in the evening. In sum, 40 videos are comprised of 20 with cap and 20 without cap, under 2 different illumination conditions. The dataset is publically available at URL: http://image.eurecom.fr/pub/mm11.

In our experiment, we tested 2 different features within 3 different ROI. The features are: 1. number of skin pixels (Skin) in ROI, 2. ratio of number of skin pixels to number of all pixels (Ratio) in ROI. Three ROI are selected automatically according to the minor diameter of ellipse returned by the head tracker. Illustrations of ROI are given in Figure 5.

Table 1. Results of the proposed system

Feature	FAR	FRR	*Classification Accuracy*
Skin_ROI1	15%	40%	72.5%
Skin_ROI2	15%	40%	72.5%
Skin_ROI3	35%	5%	80%
Ratio_ROI1	5%	30%	82.5%
Ratio_ROI2	5%	20%	**87.5%**
Ratio_ROI3	0%	25%	**87.5%**

Table 1 shows the results of our experiment. Here we define that a people without wearing cap is "accepted" and the one wearing cap is "rejected". It is clear that Ratio has better classification accuracy than Skin. Ratio_ROI2 and Ratio_ROI3 lead to the best result (up to 87.5%). For Ratio_ROI2, 1 person is wrongly "accepted" and 4 people are wrongly "rejected", whereas nobody is wrongly "accepted" using Ratio_ROI3.

Figure 5. Pre-defined ROI 1-3 (from left to right: the whole ellipse, the face region, the upper-eyebrow region)

6. CONCLUSIONS

This paper presents a solution to a newly identified occlusion problem: time-variant occlusion due to cap in entrance surveillance. The proposed system can be applied to a wide range of applications for security management in nowadays video surveillance. In particular, it can provide prior information of occlusion to face recognition system of identifying suspicious persons. It overcomes several challenges due to sensor and human behaviors in the imperfect world. In addition, it exploits the temporal information to analyze the characteristics of occlusion varying. Future works include the cap detection in crowed scenes, as well as face selection in the trade of occlusion and resolution.

7. ACKNOWLEDGMENTS

This work is partially funded by the French national project FR OSEO BIORAFALE. The authors also would like to thank Mr. Monir Azraoui for his help in simulations.

8. REFERENCES

[1] Martinez, A. M. 2002. Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 6 (June 2002), 748-763. DOI=10.1109/TPAMI.2002.1008382

[2] Kim, J., Choi, J., Yi, J., and Turk., M. 2005. Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 12 (December 2005), 1977-1981. DOI=10.1109/TPAMI.2005.242

[3] Rama, A., Tarres, F., Goldmann, L., and Sikora, T. 2008 More robust face recognition by considering occlusion information, In *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition* (Sept. 2008), vol., no., pp.1-6, 17-19.

[4] Min, R., Hadid, A., and Dugelay, J.-L. 2011. Improving the recognition of faces occluded by facial accessories, In *Proceedings of the 9th IEEE International Conference on Automatic Face & Gesture Recognition* (Santa Barbara, CA, USA, March 21-25, 2011)

[5] Du, S., Shehata, M., and Badawy, W. 2011 , "Hard hat detection in video sequences based on face features, motion and color information," In *Proceedings of the 3rd International Conference on Computer Research and Development* (March 2011) vol.4, no., pp.25-29, 11-13

[6] Birchfield, S. 1998. Elliptical Head Tracking Using Intensity Gradients and Color Histograms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR '98). Washington, DC, USA, 232-.

[7] Viola, P. and Jones, M. J. 2004. Robust Real-Time Face Detection. *Int. J. Comput. Vision* 57, 2 (May 2004), 137-154. DOI=10.1023/B:VISI.0000013087.49260.fb

[8] Wang, L., Tan, T., Ning, H. and Hu, W. 2003. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 12 (December 2003), 1505-1518. DOI=10.1109/TPAMI.2003.1251144

[9] Yang, Y.-H. and Levine, M. D. 1992. The background primal sketch: an approach for tracking moving objects. *Mach. Vision Appl.* 5, 1 (January 1992), 17-34. DOI=10.1007/BF01213527

[10] Brown, J. C., and Miller, P. J. O. 2007 Automatic classification of killer whale vocalizations using dynamic time warping. *J. Acoust. Soc. Amer.* v122. 1201-1207.

[11] Chai, D. and Ngan, K N. 1999 Face Segmentation Using Skin Color Map in Videophone Applications, *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551-564, 1999.