Pattern Recognition 44 (2011) 1068–1075

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Online learning from local features for video-based face recognition

Ajmal Mian

School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

ARTICLE INFO

Article history:
Received 21 July 2009
Received in revised form
20 August 2010
Accepted 1 December 2010

Keywords:
Online learning
Face recognition
Video-based face recognition
Local features
Clustering

ABSTRACT

This paper presents an online learning approach to video-based face recognition that does not make any assumptions about the pose, expressions or prior localization of facial landmarks. Learning is performed online while the subject is imaged and gives near realtime feedback on the learning status. Face images are automatically clustered based on the similarity of their local features. The learning process continues until the clusters have a required minimum number of faces and the distance of the farthest face from its cluster mean is below a threshold. A voting algorithm is employed to pick the representative features of each cluster. Local features are extracted from arbitrary keypoints on faces as opposed to pre-defined landmarks and the algorithm is inherently robust to large scale pose variations and occlusions. During recognition, video frames of a probe are sequentially matched to the clusters of all individuals in the gallery and its identity is decided on the basis of best temporally cohesive cluster matches. Online experiments (using live video) were performed on a database of 50 enrolled subjects and another 22 unseen impostors. The proposed algorithm achieved a recognition rate of 97.8% and a verification rate of 100% at a false accept rate of 0.0014. For comparison, experiments were also performed using the Honda/UCSD database and 99.5% recognition rate was achieved.

1. Introduction

Various physiological biometrics (e.g. iris and fingerprints) and behavioral biometrics (e.g. voice and gait) can be used for human identification. However, biometrics which can be acquired non-intrusively, using non-contact sensors and without the knowledge of the subject are of special interest due to their potential use in security applications. The human face is one of the most attractive biometrics for this purpose because it can be continuously acquired with inexpensive equipment such as a video camera. A unique application of face recognition is continuous authentication whereby the identity of a user is continuously verified by a system while introducing minimal inconvenience. Using fingerprints or keystroke dynamics for continuous authentication can restrict the user movements. However, machine recognition of faces is extremely challenging not only because the distinctiveness of facial biometrics is comparatively low [1] but because there are a number of factors over which there is little or no control. These factors include changing illumination, pose, facial expressions, facial ornamentation and occlusions.

Initial face recognition research was based on matching single pairs of images [2,3]. However, such recognition techniques do not cope well with the above challenges. More recently, recognition from 3D facial scans has been explored by many research groups [4–8]. The main limitation of 3D face recognition lies in the 3D scanning part. Compared to cameras, 3D scanners are more expensive, have lower resolution and slower acquisition time. Although 3D scanners are continuously improving on these three factors, camera technology is also improving at a fast pace. For comprehensive surveys of face recognition from images and 3D scans, the reader is referred to [9,10].

During the past few years, many research groups have developed interest in video-based face recognition because video cameras are commonly available and provide more information compared to still cameras. Moreover, motion helps in the recognition of faces [11,9]. Early video-based face recognition algorithms were frame-based. They matched individual frames from the training and test videos, and made the decision using a voting or averaging criterion. These techniques do not fully exploit the spatiotemporal information. More promising techniques match video sequences and use temporal coherence between the query and database videos in addition to the spatial information contained in individual frames. Video-based face recognition algorithms can have three possible learning modes, namely offline batch learning, online learning and hybrid learning. In batch learning the classifier is trained offline once only [12]. In this case, learning is performed on prerecorded videos and there is no feedback mechanism during video acquisition which measures if the video samples of individuals are sufficient. Moreover, the system is not updated unless a new identity is to be added in the database. In online learning, the system is completely trained online [13], however manual labeling of identities is still required. The hybrid learning approach learns generic (or specific) face models offline in a batch mode and continuously updates them during online recognition [14].

E-mail address: ajmal@csse.uwa.edu.au

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.12.001

Many video-based face recognition algorithms assume a prior knowledge of the pose and identification of pre-defined facial landmarks [15]. Others rely on supervised learning whereby each frame is manually assigned to its corresponding pose cluster [14]. Even though batch learning is an offline process, manual labeling could be very laborious and time consuming due to the bulk (up to 25 frames/s) of video data. Unsupervised learning approaches are useful for online training of the system and can prove to be time efficient, since they do not require manual labeling, as well as memory efficient because the video frames can be discarded soon after processing.

Most video-based face recognition algorithms use the complete detected face to extract global features. Global features are sensitive to registration errors, occlusions and pose variations. Local features have proved their superiority in image and 3D face recognition algorithms. However, local features have not achieved much attention in video-based face recognition because of the excessive amount of global data already available due to the temporal dimension. Local features add complexity by introducing yet another dimension, however they are important as they are robust to occlusions and provide an additional cue for accurate pattern recognition. Sivic et al. [16] used local features for retrieving different shots of a person from a movie. However, they extracted local features from pre-defined landmarks on the face. This simplifies the dimension problem as the features from the same landmarks can be compressed by projecting them to the PCA subspace [16].

This paper proposes a fully automatic video-based face recognition algorithm which performs unsupervised online learning. The frames of an input video sequence are automatically clustered during the learning phase. The learning process provides near real-time feedback on the status of enrollment. Online learning continues until the clusters have a required minimum number of faces and the distance of the farthest face from its cluster mean is below a threshold. Distance between two faces (or frames) is measured by matching their local features extracted from unordered keypoints as opposed to pre-defined landmarks. A voting scheme is used for picking the representative features and frame from each cluster. During recognition, local features from the query frames of an unknown identity (probe face) are matched with the cluster representative features and a compound frame similarity measure is used for making the final decision. By measuring the relative distance between the best two matches and the mean of the remaining matches, a confidence value is estimated for each query frame and used to reject impostors and unreliable frames of legitimate users.

This paper is an extension of the unsupervised batch learning algorithm proposed in [17]. In contrast to [17], in this paper, learning is performed online and a feedback mechanism is proposed which automatically stops the learning process only after the subject is successfully enrolled. Learning is performed in near realtime at over 12 frames/s. This paper also performs online recognition experiments on live video of subjects as opposed to [17] where all experiments were performed offline on prerecorded videos. The number of participating subjects is also increased from 20 to 72 in this paper and 22 unseen subjects are tested as impostors to calculate the ROC curves of the algorithm. Finally, automatic face detection and tracking are also integrated into the system to make the learning and recognition process fully automatic. With a database (gallery) of 50 subjects and 10 clusters per subject, the proposed algorithm updates the recognition decision every 900 ms on the average.

The rest of this paper is organized as follows. Section 2 gives a brief literature survey of video-based face recognition algorithms. Section 3 describes the local features used for face matching. Section 4 gives details of the learning and face matching algorithms. The online face recognition algorithm is described in Section 5. Section 6 describes our experimental setups and reports the recognition and verification results. Section 7 gives discussion and Section 8 concludes the paper.

2. Related work

Hadid and Pietikäinen [18] used an extended set of volume based LBP (local binary pattern) features (EVLBP) for video-based face and gender recognition. They trained an AdaBoost classifier [19] for selecting the most discriminating EVLBP features to perform fast and accurate recognition. However, a drawback of treating video as volumetric data is that it is sensitive to the face cropping (scale and location) and the frame rate. Lee et al. [12] represented each identity by ten 3D local PCA subspaces and a transition probability matrix that contained the probability of transition from one subspace (corresponding to a specific pose) to the other. They provided a complete video-based face tracking and recognition solution. The face tracker was quantitatively compared to other techniques and achieved a very accurate localization of the face. Consequently, their global features based face recognition also achieved good results.

Koh et al. [20] proposed an integrated face detection and recognition algorithm. They detected the face and defined a radial grid mapping centered on the nose to extract a feature and used a neural network for classification. The feature vector was extracted using a log-polar grid while sampling the inner nose region more densely compared to the outer regions. The authors argue that this feature vector is more robust to facial expressions. Raytchev and Murase [21] used a graph based approach for video based face recognition. They organized the video data in an incremental graph where similar views were chained together in a way similar to clustering. Their similarity measure was calculated by subtracting two images and counting the number of pixels which differ by more than a specified threshold. Such an approach will be highly dependent upon the accuracy of face tracking in scale space and will also be sensitive to illumination.

Liu et al. [13] proposed an online appearance manifold learning algorithm for video based face recognition. They proposed a method for splitting the eigenspace so that most face samples are not clustered into the same eigenspace. They also use a transition matrix similar to Lee et al. [12] to calculate the transition probabilities from one eigenspace to the other. Video-based face recognition techniques have also been designed for video indexing or shot retrieval [16,22], however, such applications are simple compared to real life face recognition problems. A common aspect about the above methods is that they all used small databases up to a maximum of 24 gallery individuals [18] and 10 impostors [20]. Compared to these techniques, we test our algorithm on a gallery of 50 individuals and 22 additional impostors.

3. Local features

The SIFT (scale invariant feature transform) [23] is used in this paper for extracting local features. However, the proposed algorithm is generic and is not tied up to specific features. An advantage of the proposed algorithm is that it does not impose any restrictions on the location of features. They can be extracted from any point on the face and need not be ordered.

SIFTs [23] are 128 dimensional unit vectors extracted at keypoints in an image. The keypoints do not conform to any specific landmarks (e.g. eye corners) on the face but are detected at the scale space extrema in the difference-of-Gaussian function convolved with the image. To qualify as a keypoint, the points must also satisfy