Journal of Signal and Image Processing
ISSN: 0976-8882 & E-ISSN: 0976-8890, Volume 3, Issue 3, 2012, pp.-114-117.
Available online at http://www.bioinfo.in/contents.php?id=48

ONLINE LEARNING FROM LOCAL FEATURES FOR VIDEO-BASED FACE RECOGNITION

KULLARWAR S.A., REDDY S. AND KANDEWAR M.

Jawaharlal Darda Institute of Engg. and Technology, Yavatmal, MS, India.
*Corresponding Author: Email- skullarwar@rediffmail.com, reddy.sarvesh007@gmail.com

Received: April 24, 2012; Accepted: May 03, 2012

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 real time 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 minimumnumber of faces and the distance of the farthest face from its cluster mean is below a threshold.
Keywords- Online learning, Face recognition, Video-based face recognition ,Local features, Clustering , face aging.

Citation: Kullarwar S.A., Reddy S. and Kandewar M. (2012) Online learning from local features for video-based face recognition. Journal of Signal and Image Processing, ISSN: 0976-8882 & E-ISSN: 0976-8890, Volume 3, Issue 3, pp.-114-117.

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. 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.

Local features

The SIFT (scale invariant feature transform) [9] 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 [7] 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 other conditions including high contrast, good localization along an edge and principal curvature ratio of above a threshold.

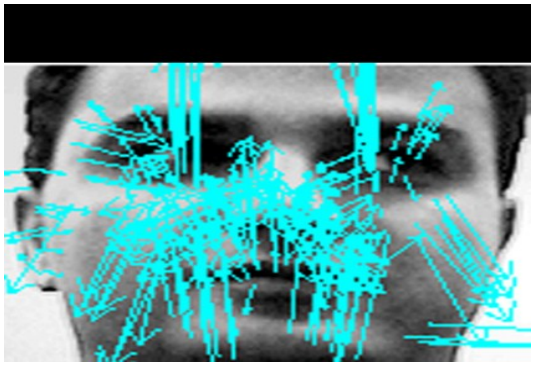

Fig. 1- Representative SIFT features of a cluster drawn on the cluster representative face. The direction of the arrow points towards the orientation and its length represents the magnitude of the gradient.

Unsupervised learning

Face matching

For a given training video sequence of an identity, every face is matched to every other face in order to construct a N × N (where N is the number of frames) similarity matrix. Since the matrix is symmetric, only matches must be performed. The similarity between two faces is determined by matching their respective SIFT features using equation

$$e = cos^{-1}(f_a(f_b)^T)$$

where fa and fb correspond to the SIFT feature vectors from face a and b respectively. The pairs of SIFTs which had the minimum error e were considered matches and only one-to-one matches were allowed. For example, if a feature in face b turned out to be the best match to more than one feature in face a, only the one with the minimum value of e was considered as its match. Moreover, a distance constraint was used to avoid matching SIFTs from far off points in the two face images. Due to these constraints, different faces ended up with a different number of SIFT matches. The overall similarity of the two faces was determined by normalizing the average error e between their matching pairs of SIFTs and the total number of matchesm. Both measures were normalized on the scale of 0 to 1 and combined using a weighted sum rule.

Frame clustering

The above matching process results in an N×N symmetric matrix of similarity measures which is used to automatically cluster the N frames. Hierarchical clustering with mean similarity distance was used in our experiments. The total number of clusters per video sequence (and hence identity) was empirically chosen to be 20 for our initial offline batch learning. This number was chosen keeping in view that there are three degrees of freedom in pose, five common facial expression types and that the variation in pose and expression can occur in many possible combinations e.g. pitch + yaw + smile. The number of clusters were later reduced to 10 for our online learning algorithm in order to achieve real time performance.

Selection of cluster representatives

One of the motivations behind clustering is data compression whereby each cluster is represented by its subset. Global features based algorithms or the ones that extract local features from predefined landmarks, generally use the mean features as representatives of clusters. However, this is not possible in the proposed algorithm as the local features are extracted from arbitrary key points as opposed to pre-defined landmarks. Therefore, a voting scheme was used to select the representative local features from each cluster as follows. In addition to the representative features, the face whose local features get the maximum accumulative votes is selected as the cluster representative. This representative face can be useful for a global features-based classifier that runs in parallel to the existing one in order to increase the accuracy of recognition. However, global features based face recognition is outside the scope of this paper.

The key point detection process of SIFT generally finds different numbers of key points in different frames, hence biasing the matching process in favor of frames with more features. This is not critical in frame clustering as all frames belong to the same video sequence and are therefore acquired in similar illumination conditions. However, this biasing could be critical during recognition when different video sequences are matched. Selecting a fixed n number of features for each cluster removes the biasing due to the number of features as well as the number of frames per cluster. In this paper, n was fixed at 200.

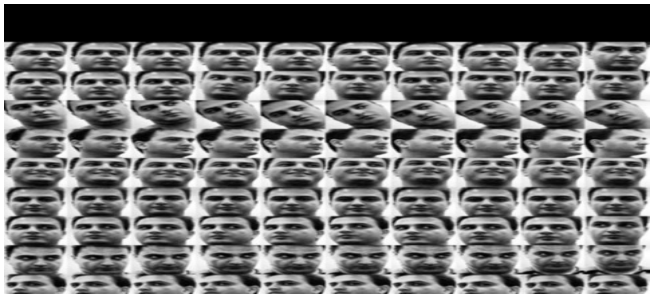

Fig. 2- Sample clustered faces from the Honda/UCSD database. Each row contains a different cluster. Notice that faces with similar facial expression are also clustered together (fifth row).

Since there is intrinsic uncertainty for face aging, we propose two criteria to evaluate the face aging results.

1) The accuracy of simulation. For each age group we select 80 real images from our data set and 80 simulated images synthesized using our algorithm. Then these images are given to 20 volunteers for age estimation. By analyzing the results with Analysis of Variance (ANOVA), we find no significant difference in age estimation performance between real images and synthetic images.
2) Preservation of the identity. We collect real aging sequences of 20 individuals from relatives and friends; for each individual, we synthesize one aging sequence from the photo at the initial age group and then 20 volunteers are asked to identify the individuals in the two sets. The ANOVA analysis of recognition results shows that our face aging model preserves face identity effectively.

Face aging

At level one, the face aging effects reflect the change of global face shape, skin color darkening, and drop of muscles. We select aging patterns based on geometric and photometric similarities.

For each face, we have 90 facial points describing the facial geometry Tgeo face;t. TPS warping energy measuring the cost for aligning two face geometries is used as a natural shape distance. The appearance distance is computed as the KL distance between histograms of corresponding filter responses (mean, variance, etc.) of two aligned faces. As studied in [1], [3] there occur certain noticeable bony and soft tissue changes in shape, size, and configuration during adult aging, and the shape changes in muscular regions is larger than in bony regions. We compute the differences between mean face shapes of different age groups as adopt the mean shape changes as soft constraints during warping of face shape as age increases.

Improved poisson image editing

As explained in Sections II-A and B, we conduct image fusion in two aspects separately. Fusion in the non sketchable region generates low-resolution result, while fusion in the sketchable region synthesizes high-resolution details. We introduce image-editing techniques to merge the synthetic low-resolution face image Il and high-resolution face image Ih seamlessly to generate the final fusion result.

Matching Non-Corresponding Face Regions via Canonical Correlation Analysis

In this section, we first describe the feature extraction method utilized in the proposed approach. Then we describe
the canonical correlation analysis and its application to matching non-corresponding face regions.

Face Feature Extraction and Corresponding

Region Matching Low-level visual feature extraction is the first and very important step in face recognition. Since local Gabor feature is successful and widely used in face recognition, we applied it for visual feature extraction. Similar to the approach in [5], we extract Gabor magnitude features of 5 scales and 8 orientations from the face images.

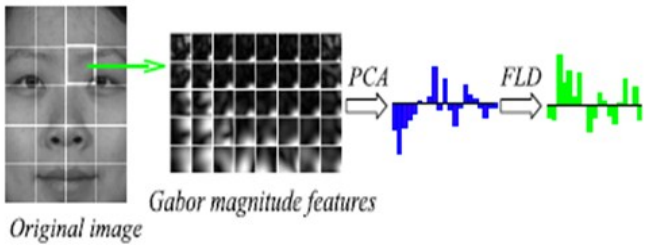

Fig. 3- The patch division and process of feature extraction.

As shown in Figure 3, we divide face images into several non-overlapping regions for illustrating the relationship between different parts of face. Gabor features are sampled from each region. To enhance the representation power and
reduce the dimension of feature vector principle component analysis(PCA)[6] is performed on each patch. After that
we perform Fishers linear discriminant analysis (FLD)[12] for further enhancement. For matching corresponding face regions, the similarity between two faces (x; y) can be simply computed by summing the similarities of each corresponding patch pair. Using correlation as the metric, the total similarity Stotal is given by

$$S_{total} = \frac{1}{m}\sum_{i=1}^{m}\frac{\langle x_i, y_i\rangle}{\|x_i\| \cdot \|y_i\|},$$

where the *⟨a; b⟩* denotes the dot product of a and b.

Matching Non Corresponding Face Regions via Canonical Correlation Analysis

In this paper we propose a novel method in which faces are recognized by matching non-corresponding regions. The method include two steps, i.e., the training step and the testing step respectively. First, we construct a series of coupled training sets, each for one pair of non-corresponding regions, by sampling one pair of patches on each face image. Then, CCA is performed on each training set for learning two sets of basis vectors. Each set of vectors correspond to one face region. In the testing phase, patch vectors are projected onto corresponding basis vectors. Similarity between two non-corresponding patches is measured by comparing their projections. In other words, the basis vectors obtained by CCA can transform vectors of two non-corresponding regions into a unified latent subspace. Non-corresponding region matching is performed in this latent subspace.

Online learning

Our initial experiments were performed using unsupervised batch learning mode using the Honda/UCSD database [2]. All the frames of the training video were matched to generate an N×N symmetric similarity matrix which was used to cluster the frames select the representative features. This approach is inefficient for large training videos (because of the O(N2) complexity) and does not provide any feedback on the learning outcome. To overcome these limitations, we propose an online unsupervised algorithm which provides near real time feedback on the learning process. During online learning, the face is automatically detected, tracked and cropped from the input video frames of a subject. Each face is matched with all the previously acquired faces to construct an N×N similarity matrix. An ith frame is matched with all frames and each similarity score is recorded at two symmetric locations in the similarity matrix. The votes for feature matches required for selecting cluster representatives are simultaneously casted during matching. To make real time learning possible, the algorithm was implemented in C++ and multi-threading was used while keeping N fixed at 100. After constructing a 100×100 similarity matrix, the frames are clustered and the number of clusters with less than a predefined number of faces are counted. If this count is above a threshold or the distance of the farthest face from its cluster mean is below a threshold, the learning process is considered complete. Otherwise, the last m faces with the maximum distance from their respective cluster means are discarded along with their features and replaced with another m freshly acquired faces. These faces are matched to the remaining faces (and among themselves as well) to update the similarity matrix and the clustering is repeated. This process continues until the above two learning criteria are met. The learning criteria are meant to avoid uneven distribution of faces in the clusters e.g. if a subject remains still throughout the online learning process, most of the faces will go into one or a few clusters . In other words, most of the data will be redundant. However, it is desirable to capture different poses and perhaps differ-

ent expressions of a face during the learning process. In an unsupervised learning approach, the easiest way to achieve this is to ensure that the faces are evenly distributed among the clusters . The two thresholds for the learning criteria are a trade off between user (subject) convenience and the learning effectiveness. On one extreme, if only a few clusters are required to have more than the minimum faces, the learning algorithm will stop after one iteration after acquiring N frames. On the other extreme, if all clusters are required to have exactly the same number of frames, the algorithm will take too many iterations and, depending upon the dynamics of the subject, may never stop. This will be very inconvenient to the subject. The second criterion is used to detect and remove erratic frames which occur as a result of inaccurate face cropping, motion blur, occlusion etc.

Effects of identity changes

In real world scenarios, identities do not abruptly change in video sequences. Especially when a face is tracked, the tracking sequence itself gives significant information when a change in identity occurs, i.e. when one face leaves and another face enters the field of view of a camera. Studying the effects of identity changes without using face tracking information would be hypothetical. However, for the interest of analyzing the proposed face recognition algorithm as opposed to face tracking, we will only consider the effects on similarity scores when identities abruptly change in a video sequence. The compound temporal similarity score gets biased towards an identity with the passage of time which is likely to cause a lag in the recognition process when there is a change in identity in the test video. The batch temporal recognition will also have a lag but it will be limited to about 0.5f frames.

Online Face Recognition In Live Video

For our online learning and recognition experiments, the number of identities in the gallery was increased to 50. All participating individuals were encouraged to change their pose and expressions in any order they preferred during the training as well as the recognition sessions. There was a two month gap between the training and test sessions. Moreover, an additional 22 individuals were tested as impostors, i.e. they were neither included in the gallery and nor appeared in the training process. In these experiments, live video was acquired for the participating subjects and the video frames were discarded soon after they were matched. Only the cluster representative features were retained during training and the similarity scores were retained during recognition. A recording of the video was not available to repeat the recognition offline. This is a realistic scenario and Is much more challenging than recognition on pre-recorded videos. Our un optimized implementation of the algorithm in C++ using a 2.4 GHz quad core machine could perform about 550matches per second. Since there were 50 identities in the gallery and 10 clusters per identity, the recognition results were updated every 900ms and displayed on top of each frame. Moreover, the similarity scores were saved in a log file for analysis. snapshots of our video-based face recognition algorithm. Notice that the ellipse,formed by CAMSHIFT tracking, changes in size when the pose changes and does not perfectly enclose the face. However, our algorithm is robust to these problems as it extracts scale invariant features from keypoints on the face which do not depend upon the cropping window. the relative performance of batch temporal recognition and compound temporal recognition. Compound temporal recognition performs better again and reaches a peak identification rate of 97.8% at 13 frames. These results are similar to the ones obtained on the UCSD database where a maximum identification rate of 99.5% was reached at the rank recognition rate of the compound temporal recognition for different number of frames. A 100% identification rate is reached at rank 3 using 13 frames and at rank 7 using 9 frames

Conclusion

This paper presented an online learning algorithm for video based face recognition. Learning is performed online at over12 frames/s and a feedback mechanism is used to ensure that sufficient pose variations of the subject are enrolled. The proposed algorithm is robust to large scale pose and expression variation sand can handle partial occlusions. It is robust to errors in face tracking as it extracts scale invariant features from arbitrary key points which are independent of the location and scale of the cropping window. Local features of a query face are matched with the database and a compound temporal similarity measure is used to establish its identity. Recognition experiments were performed on a standard Honda/UCSD database and 99.5% recognition rate was achieved. Online learning and recognition experiments were also performed on live subjects. With a database of 50 enrolled subjects and another 22 unseen impostors, the algorithm achieved a recognition rate of 97.8% and a verification rate of 100% at 0.004 FAR.

References

[1] AjmalMian, *Online learning from local features for video-based face recognition.*

[2] Mian A. (2008) *IEEE Face and Gesture Recognition*.

[3] Lowe D. (2004) *International Journal of Computer Vision* 60 (2), 91-110.

[4] Lee K., Kriegman D. (2005) *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 852-859.

[5] Faltemier T., Bowyer K., Flynn P. (2008) *IEEE Transactions on Information Forensics and Security,* 3 (1), 62-73.

[6] Lu X., Jain A.K., Colbry D. (2006) *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28 (1), 31-43.

[7] Mian A., Bennamoun M., Owens R. (2007) *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 29 (11), 1927-1943.

[8] Passalis G., Kakadiaris I., Theoharis T. (2007) *ACM Computing Survey,* 29 (2), 218-229.

[9] Zhao W., Chellappa R., Phillips P.J., Rosenfeld A. (2003) *ACM Computing Survey,* 35 (4), 399-458.

[10] Bowyer K., Chang K., Flynn P. (2006) *Computer Vision and Image Understanding,* 101 (1), 1-15.

[11] Ahonen T., Hadid A. and Pietikainen M. (2006) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (12), 2037-2041.

[12] Belhumeur P.N., Hespanha P. and Kriegman D.J. (1997) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (7), 711-720.