Video-Based Face Recognition Using A Metric of Average Euclidean Distance

Jiangwei Li, Yunhong Wang, Tieniu Tan

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100080 P.R.China
{jwli, wangyh, tnt}@nlpr.ia.ac.cn

Abstract. This paper presents a novel approach for video-based face recognition. We define a metric based on an average L_2 Euclidean distance between two videos as the classifier. This metric makes use of Earth Mover's Distance (EMD) as the underlying similarity measurement between videos. Earth Mover's Distance is a recently proposed metric for geometric pattern matching and it reflects the average ground distance between two distributions. Under the framework of EMD, each video is modeled as a video signature and Euclidean distance is selected as the ground distance of EMD. Since clustering algorithm is employed, video signature can well represent the overall data distribution of faces in video. Experimental results demonstrate the superior performance of our algorithm.

1 Introduction

Face recognition based on video has been a focus recently [1-6]. It is very useful in the application of video surveillance and access control. Compared to still-based face recognition technology, multiple frames and temporal information facilitate face recognition. The discriminative information can be integrated across the video sequences. However, poor video quality, large illumination and pose variations, partial occlusion and small size images are the disadvantages of video-based face recognition. To overcome the above problems, many approaches, which attempt to utilize multiple frames and temporal information in video, are proposed. Based on whether the temporal information is utilized or not, these schemes can be divided into sequential approach and batch approach.

Sequential approach assumes temporal continuity between two adjacent samples. The continuity property propagates face position and identity frame by frame. The previous tracking and recognition result can be utilized for current face tasks. Zhou [2] proposes a tracking-and-recognition approach, which utilizes a very powerful unified probabilistic framework to resolve uncertainties in tracking and recognition simultaneously. Lee [3] represents each person with an appearance manifolds expressed as a collection of pose manifolds. In recognition, the probability of the test image from a particular pose manifold and the transition probability from the previous frame to the current pose manifold are integrated. Liu [4] applies adaptive HMM to perform video-based face recognition task.

The other is batch approach, which assumes independence between any two samples, thus the dynamics of image sequences is ignored. It is particularly useful to recognize a person from sparse observations. The main idea of batch approach is to compute the similarity function $f(A,B)$, where A and B are training and testing video, respectively. The greater value of $f(A,B)$ indicates A and B are more likely sampled from the same individual. The way to define $f(A,B)$ differentiates various batch methods [5,6].

In this paper, we propose a novel model to identify the querying video. It is based on the measurement of average Euclidean distance between two videos so it is one of batch approaches. Instead of modeling set of video images as subspace [5] or Gaussian distribution [6], we represent the distribution of each set with a video signature. Video signature reflects the complex distribution of video data in image space. Earth Mover's Distance (EMD) is the proposed metric for average distribution distance measurement between two signatures. For simpleness and effectiveness, Euclidean distance is suggested to be the underlying ground distance of EMD. A new similarity function based on the average Euclidean distance metric is established and we verify its performance on a combined database.

This paper is organized as follows. Section 2 gives a brief review of some related work. In Section 3, the metric of average Euclidean distance is introduced. Section 4 discusses experimental results. At last, we conclude the paper and prospect future work.

2 Related Work

As mentioned above, for batch approach of video-based face recognition, the purpose is to define the similarity function $f(A,B)$, where A and B are training and testing video, respectively. MSM [5] defines similarity function as follows:

$$f(A,B) = \cos(\theta) = \max \frac{u_A{}^T u_B}{\|u_A\| \cdot \|u_B\|} \tag{1}$$

u_A and u_B are eigenvectors of A and B, respectively. MSM is thought that some discriminative statistical characteristics, e.g., eigenvalues or means of the data, are not considered. For K-L divergence [6] method, it is defined as:

$$f(A,B) = -\frac{1}{2} \cdot (\log(\frac{|\Sigma_A|}{|\Sigma_B|}) + tr(\Sigma_B \Sigma_A^{-1} + \Sigma_A^{-1}(m_A - m_B)(m_A - m_B)^T) - d) \tag{2}$$

In this formula, Σ is the covariance matrix, while m is the mean vector. d corresponds to the dimensionality. It assumes each set of video images can be modeled as a multivariate Gaussian distribution, which is not precise enough to be the underlying distribution due to multiple poses and expressions in video. In addition, the computation of Equation (2) is very time-consuming since Σ is always a singular matrix [8].

3 The Proposed Metric

In this paper, to define the similarity function $f(A,B)$ between two videos, we introduce the notion of Earth Mover's Distance (EMD). EMD is a recently proposed metric for geometric pattern matching. It compares two distributions that have the same weights and it is proved much better than some well-known metrics (e.g., Euclidean distance between two vectors). It is based on an optimization method for the transportation problem [7] and is applied to image retrieval [9,14] and graph matching [11,12]. The name is suggested for road design [13].

3.1 Video Signatures

Given a set of points in image space, we represent the set with a signature. The signature is composed of numbers of clusters of similar features in a L_2 space. Each cluster is attached to a weight, which reflects the ratio of the number of images in this cluster to the total number of images. For video-based face recognition, each video corresponds to a distribution of points in image space and can be modeled as a ***video signatur***e. We employ the technology of vector quantization [10] for clustering since it performs efficiently. Each cluster contributes a pair (u , p_u), where u is the prototype vector of the cluster and p_u is its weight which is the fraction of face images in the cluster.

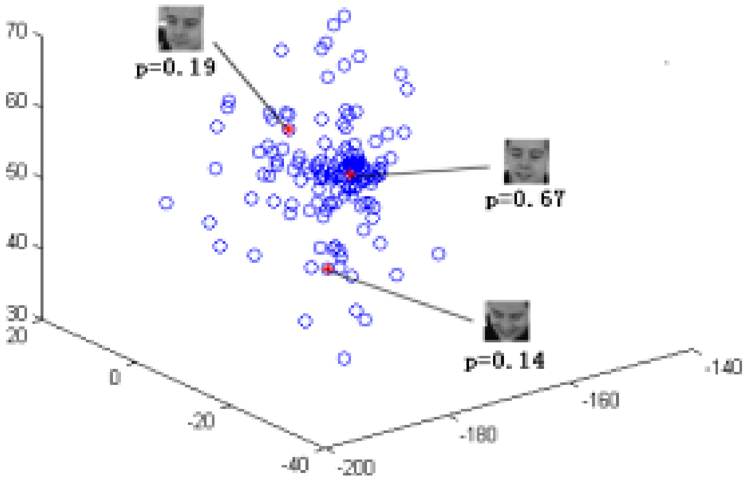

Fig. 1. A signature for video-based face recognition

For videos, poses and expressions change constantly. The images in a video form a complex distribution in high dimensional image space. It can not be simply expressed by a single subspace or a single multivariate Gaussian model. Since clustering algorithm is used, signature can well represent the overall data distribution of video data. Each cluster corresponds to a pose manifold. In addition, with clustering, some degree of variations, e.g., illumination, poses and expressions, can be tolerated. Moreover, changing the number of clusters, it provides a compact and flexible method to represent data distribution. More clusters are used, more precise the model is. Fig. 1 is an example of a signature in a reduced dimensionality space. Each signature contains a set of prototype vectors and their corresponding weights. In Fig. 1, the prototype is labeled with a red “＊” and the weight is denoted under the corresponding image.

3.2 Average Euclidean Distance Between Video Signatures

Assume two videos A and B are modeled as video signatures. We can imagine A is a mass of earth, and B is a collection of holes. EMD is a measurement of the minimal work needed to fill the holes with earth. This is the reason why it is named “Earth Mover’s Distance”. Fig. 2 shows an example with three piles of earth and two holes.

EMD can be formalized as the following linear programming problem: Let $A=\{(u_1, p_{u_1}),\ldots,(u_m, p_{u_m})\}$ and $B=\{(v_1, p_{v_1}),\ldots,(v_m, p_{v_n})\}$, where u_i, v_j are the prototype vectors of clusters of A and B, respectively, and p_{u_i} and p_{v_j} are their corresponding weights. The cost to move an element u_i, to a new position v_j is the cost coefficient c_{ij}, multiplied by d_{ij}, where c_{ij} corresponds to the portion of the weight to be moved, and d_{ij} is the ground distance between u_i and v_j. EMD is the sum of cost of moving the weights of the elements of A to those of B. Thus the solution to EMD is to find a set of cost coefficients c_{ij} to minimize the following function:

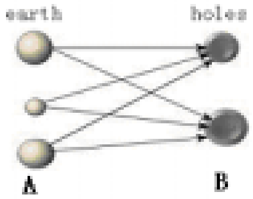

Fig. 2. An example of EMD

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} d_{ij} \tag{3}$$

subject to: (i) $c_{ij} \geq 0$, (ii) $\sum_{i=1}^{m} c_{ij} \leq p_{v_j}$, (iii) $\sum_{j=1}^{n} c_{ij} \leq p_{u_i}$, and (iv) $\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} = \min(\sum_{i=1}^{m} p_{u_i}, \sum_{j=1}^{n} p_{v_j})$. Constraint (i) indicates only positive quantity of "earth" is allowed to move. Constraint (ii) limits the quantity of earth filled to a "hole". Each hole is at most filled up all the capacity. Constraint (iii) limits the quantity of earth provided to holes. Each pile of earth provides at most its capacity. Constraint (iv) prescribes that at least one signature contributes all its weights. If the optimization is successful, then EMD can be normalized as:

$$EMD\ (A, B) = \frac{\min(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} d_{ij})}{\min(\sum_{i=1}^{m} p_{u_i}, \sum_{j=1}^{n} p_{v_j})} \tag{4}$$

EMD extends the distance between single points to the distance between sets of points and it reflects the average ground distance that weights travels according to an optimal flow. In general, the ground distance d_{ij} can be any distance and it will be chosen according to the problem we encounter. For the simpleness and effectiveness, Euclidean distance is proposed to be the underlying ground distance of EMD. Since EMD is the basic distance framework to represent the average ground distance between signatures and Euclidean distance is the underlying ground distance, it is named as "a metric of average Euclidean distance". Based on this metric, the similarity function between the querying video A and the reference video B can be defined as:

$$f(A, B) = \exp(-\frac{EMD(A, B)}{\sigma^2}) \tag{5}$$

where σ is a constant for normalization. The value of the function shows the degree of similarity between A and B.

Particularly, if some weights of clusters are smaller than a threshold, we discard these clusters since it contributes a little for matching. For videos, these clusters generally consist of faces on bad condition, which deviate far away from normal face clusters. EMD provides a natural solution to this kind of partial matching. However, EMD with partial matching is not a metric for the distance measure of two distributions.

4 Experimental Results

4.1 Experimental Database

We use a combined database to evaluate the performance of our algorithm. The database can be divided into two parts: (i) Mobo (Motion of Body) database. Mobo database was collected at the Carnegie Mellon University for human identification. There are 25 individuals in the database. (ii) Our collected database. This part is collected from advertisements, MTV and personal videos. There are 15 subjects in the database. Totally, our combined database contains 40 subjects, and each subject has 300 face images. Fig. 3 shows some faces cropped from sequences in the database. Using the very coarse positions of eyes, we normalize it to 30×30 pixels and use it for experiments. Some location errors, various poses and expressions can be observed in the database.

4.2 Experimental Scheme

In order to verify the performance of our proposed algorithm, the scheme of data partition is illuminated as follows: given a video in database, we select the first 100 frames for training. In the remaining 200 frames, we randomly select a starting frame K and a length M. Then frames $\{I_{K+1}, I_{K+1,} \cdots, I_{K+M}\}$ form a sequence for testing. This is similar to the practical situation where anyone can come to the recognition system at any time with any duration [4].

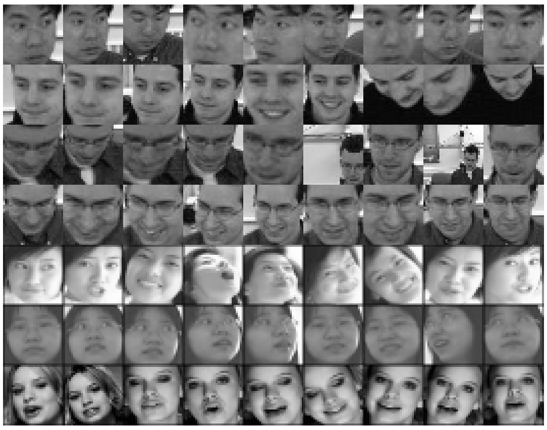

Fig. 3. Some cropped faces from sequences

Two experiments are performed. The first experiment changes the number of clusters of video signature. It wants to disclose how many clusters in signature are most beneficial to face recognition. The second experiment compares the algorithm with other general batch methods to demonstrate its performance.

4.3 Recognition Rate Vs. Number of Clusters

For video-based face recognition, each video forms a video signature in image space by clustering algorithm. We think since clustering algorithm is used, signature can well represent the overall data distribution of videos. We further deduce that the number of clusters in a signature may affect the recognition rate. Based on above data partition scheme, testing video is obtained with a random starting frame and a random length. This experiment is performed four times with different length of testing video. The purpose of the experiment is to disclose the relationship between recognition rate and the number of clusters in a signature and the length of testing video. Fig. 4 illustrates the experimental results.

In Fig.4, the horizontal axis represents the number of clusters used in a video signature. The vertical axis denotes the recognition rate. The legend on left top corner represents how many frames are used for testing. The length of testing video is 51, 92, 120 and 160, respectively. The number of clusters changes from one to eight. When only a cluster is used, EMD is actually the Euclidean distance between centers of videos. Directly using center vector to represent a video is very coarse so that the recognition rate is only about 55%. With the increment of the number of clusters, the model of video signature becomes more and more precise. When more than four clusters are used, the recognition rate is nearly 100%. From this figure, we can note that no matter how long the testing video is, five clusters is a preferable choice for recognition. This experiment demonstrates that video signature can well represent the overall data distribution of sets. Furthermore, since each cluster may correspond to a pose manifold of complex distributed video and EMD reflects the average Euclidean distance between videos, it is a reasonable result of high recognition rate.

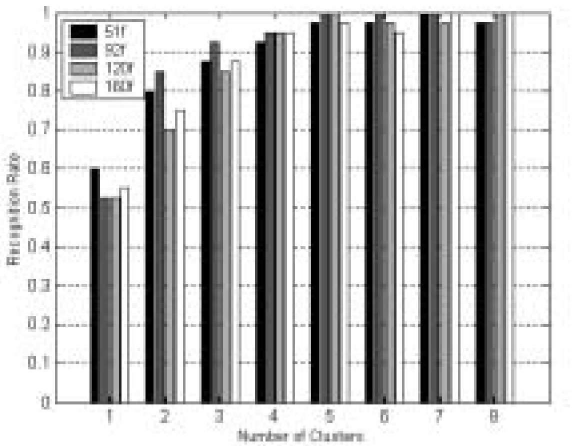

Fig. 4. Recognition rate Vs. Number of clusters

4.4 Comparison With Other Algorithms

In this experiment, three batch algorithms are assembled together to compare their performance. They are MSM (Mutual Subspace Method), KLD (K-L Divergence) and

EMD (Earth Mover's Distance). The underlying data distribution model of MSM is a single subspace, while that of KLD is a single Gaussian model and EMD is a video signature. The experiment is done ten times to obtain the recognition rate curve as shown in Fig. 5.

In Fig. 5, the horizontal axis shows the length of testing video. It ranges from 19 to 189 frames. The vertical axis shows the recognition rate. For MSM, we use all eigenvectors to compute the similarity function. For EMD, five clusters are contained in a video signature based on the former experiment. From this figure, we can observe that when less than 20 frames are for test, the performance of KLD is very weak. It is because that a representative Gaussian model needs more training samples. Its performance becomes better with the increasing of testing frames. However, the performance of KLD is still the worst one. This phenomenon demonstrates that K-L divergence between probabilistic models is not an effective classifier and a single Gaussian model is not robust for expressing complex data distribution of video. We also note that the performance of MSM and EMD is similar. Their difference is that when the length of testing video is short, EMD is much better than MSM, especially in the case that only 19 frames are used for testing. It proves that a single subspace can not reflect the distribution of small quantity of data. Video signature is a much more reasonable model of data distribution and EMD is a robust metric for classification since it reflects the average Euclidean distance between video signatures.

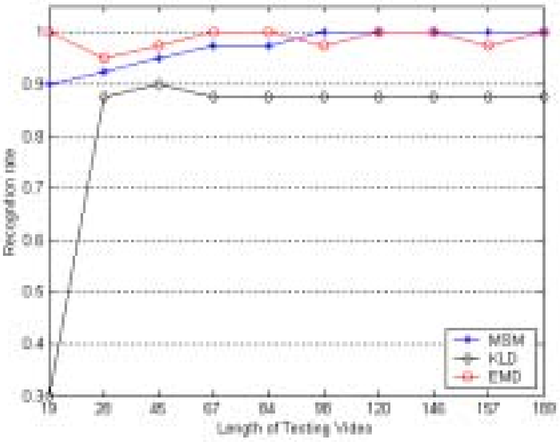

Fig. 5. Comparison of MSM, KLD And EMD

5 Conclusion

For video-based face recognition, conventional batch approaches supply two classical methods to estimate the similarity between testing video and training video. The one is to compute the angle between subspaces [5], and the other is to find K-L divergence [6] between probabilistic models. In this paper, we consider a most straightforward method of using distance for matching. We propose a metric based on an average L_2

Euclidean distance between two videos as the classifier. This metric is established based on Earth Mover's distance and Euclidean distance is suggested to be the underlying ground distance metric. Signatures are built for modeling data distribution of image sets. This model is much better than subspace model [5] and single probabilistic model [6] since it divides a video into several clusters corresponding to different poses. In future, we will consider the update method to improve the representative capability of signatures. Moreover, as in [4], time information and transformation probability will be considered to build a more reasonable model to represent a video.

References

[1] W.Zhao, R.Chellappa, A. Rosenfeld and P.J Phillips, "Face Recognition: A Literature Survey", Technical Reports of Computer Vision Laboratory of University of Maryland, 2000.

[2] S. Zhou and R.Chellappa, "Probabilistic Human Recognition from Video", In Proceedings of the European Conference On Computer Vision, 2002.

[3] K.C.Lee, J.Ho, M.H.Yang, D.Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds", In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2003.

[4] X.Liu and T.Chen, "Video-Based Face Recognition Using Adaptive Hidden Markov Models", In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2003.

[5] O.Yamaguchi, K.Fukui, K.Maeda, "Face Recognition using Temporal Image Sequence," In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998.

[6] G.Shakhnarovich, J.W.Fisher, T.Darrell, "Face recognition from long-term observations", In Proceedings of the European Conference On Computer Vision, 2002.

[7] G. B. Dantzig, "Application of the simplex method to a transportation problem", In Activity Analysis of Production and Allocation, 359–373, John Wiley and Sons, 1951.

[8] B. Moghaddam, A. Pentland, "Probabilistic visual learning for object representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997.

[9] Y.Rubner, C.Tomasi, L.J.Guibas, "Adaptive Color-Image Embedding for Database Navigation", In Proceedings of the Asian Conference on Computer Vision, 1998.

[10] "Learning Vector Quantization (LVQ)", < http://www.willamette.edu/~gorr/classes/cs449/Unsupervised/competitive.html>.

[11] Y.Keselman, A.Shokoufandeh, M.F.Demirci, S.Dickinson, "Many-to-Many Graph Matching via Metric Embedding", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[12] M.F.Demirci, A.Shokoufandeh, Y.Keselman, S.Dickinson, L.Bretzner, "Many-to-Many Feature Matching Using Spherical Coding of Directed Graphs", In Proceedings of the 8th European Conference on Computer Vision, 2004.

[13] J.Stolfi, "Personal Communication", 1994.

[14] S.Cohen, L.Guibas, "The Earth Mover's Distance under Transformation Sets", In Proceedings of the 7th IEEE International Conference On Computer Vision, 1999.