# Video-based Face Recognition Using Earth Mover's Distance

Jiangwei Li[1], Yunhong Wang[1,2], and Tieniu Tan[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, P.R.China
[2] School of Computer Science and Engineering, Bei Hang University, Beijing, 100083, P.R.China

**Abstract.** In this paper, we present a novel approach of using Earth Mover's Distance for video-based face recognition. General methods can be classified into sequential approach and batch approach. Batch approach is to compute a similarity function between two videos. There are two classical batch methods. The one is to compute the angle between subspaces, and the other is to find K-L divergence between probabilistic models. This paper considers a most straightforward method of using distance for matching. We propose a metric based on an average Euclidean distance between two videos as the classifier. This metric makes use of Earth Mover's Distance (EMD) as the underlying similarity measurement between two distributions of face images. To make the algorithm more effective, dimensionality reduction is needed. Fisher's Linear Discriminant analysis (FLDA) is used for linear transformation and making each class more separable. The set of features is then compressed with a signature, which is composed of numbers of points and their corresponding weights. During matching, the distance between two signatures is computed by EMD. Experimental results demonstrate the efficiency of EMD for video-based face recognition.

## 1 Introduction

Recently, more and more researchers are focusing on face recognition from video sequences [1][2][3][4][5][6], which is very useful in applications of surveillance and access control. Compared to still-based face recognition technologies, multiple frames and temporal information facilitate the process of 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 image are the disadvantages of video-based face recognition. To overcome 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 are ignored. It is particularly useful to recognize a person from sparse observations. The main idea of batch approach is to compute the similarity between two videos. For instance, Mutual Subspace Method (MSM)[5] defines the similarity by the angle between two subspaces spanned by the basis of image sets. Shakhnarovich [6] used multivariate Gaussian models to represent the densities of face sets, and K-L divergence between models is used for matching.

The main problems of the above batch methods are heavy computational cost and not precise models. It is not efficient to estimate the subspace or Gaussian model directly in image space. Moreover, they are not considering the complex data distribution of video data. Both of the subspace and the Gaussian model are only effective to the convex data sets. But in video, head poses, face expressions and illumination change constantly, the shape of data distribution is largely non-convex, more robust model is needed.

Our algorithm is a novel method of batch approaches. In the paper, instead of modeling the data distribution directly in high dimensional image space, we firstly reduce the dimensionality. There we use Fisher's Linear Discriminate (FLDA)[7] to map sets of images to groups of points in low-dimensional feature space. With a linear transformation, FLDA makes sets of images more compact and separable. Furthermore, it reduces the computational consuming. Each video yields a set of points in feature space. We consider a more reasonable model to estimate the distribution of each set. The match of videos can be viewed as the geometric match of sets in feature space. We use the conception of signature to represent each set. By clustering algorithm, the points in a set are grouped into several clusters. The signature is composed of means and weights of these clusters. In fact, it reflects complex data distribution of the set in feature space. So the match problem turns to be the distance measurement of two signatures. Earth Mover's Distance (EMD) is proposed for this purpose. EMD is based on an optimization method for the transportation problem[8]. It computes the minimum work done by moving the weights of one signature to another. EMD is good metric for the comparison of two distributions and in addition, it is adaptive for partial matching, since some faces with large pose variations are thought to be useless and should be discarded in matching. However, when partial matching, EMD is not a metric. In our method, with FLDA for linear transformation, face images are well represented and the computational cost becomes low. In addition,

each distribution of observations in feature space can be efficiently modeled as a signature and the similarity of two videos can be easily and accurately estimated by EMD.

## 2 Earth Mover's Distance for Recognition

Earth Mover's Distance is a general metric to compare two distributions that have the same weights. To accommodate pairs of distributions that are "not rigidly embedded"[12], the definition of EMD is:

$$EMD(A, B) = min_{f \in F} EMD(A, f(B)) \tag{1}$$

where $A$ and $B$ are two distributions. The purpose of this equation is to seek a transformation $f$ that minimizes $EMD(A, B)$ . "FT iteration"[12] is proposed to the solution of object function $f$. In this paper, considering its application to video-based face recognition, we define it as:

$$EMD(A, B) = \max_{g \in G} EMD(g(A), g(B)) \tag{2}$$

where $g$ is a linear transformation to project two distributions onto feature space so as to maximize $EMD(A, B)$.

### 2.1 Linear Transformation

As mentioned above, considering the efficiency, the techniques of linear subspace, e.g., PCA[10], FLDA[7] and ICA[11], are taken into account. For simplicity and validity, we use FLDA. In FLDA, between-class matrix $S_b$ and within-class matrix $S_w$ are defined as:

$$S_b = \sum_{i=1}^{H} N_i (m_i - m)(m_i - m)^T \tag{3}$$

$$S_w = \sum_{i=1}^{H} \sum_{x_k \in L_i} (x_k - m_i)(x_k - m_i)^T \tag{4}$$

where $m_i$ is the mean of the image set $L_i$ , and $N_i$ is the number of images in $L_i$.

The linear transformation matrix $W$ maximizes the following optimal criterion:

$$W = argmax_{\phi} \frac{|\Phi^T S_b \Phi|}{|\Phi^T S_w \Phi|} \tag{5}$$

For video-based face recognition, $S_w$ is generally a full rank matrix. So $W$ can be obtained by seeking the eigenvectors of $S_w^{-1} S_b$ directly.

Using linear transformation, the dimensionality of video data is much reduced. It preliminary solves the problem of heavy computation for video-base face recognition. Furthermore, with FLDA, some variations contained in video are modelled, so the sets of images become compact and separable.

### 2.2 Earth Mover's Distance

After linear transformation, we obtain two feature distributions $g(A)$ and $g(B)$. In order to define the similarity function $f(A, B)$ between two videos, we introduce the notion of Earth Mover's Distance (EMD). EMD is a general distance measure with application to image retrieval[12][13] and graph matching [16][17]. It is proved much better than other well-known metrics (e.g., Euclidean distance between two vectors). The name is suggested by Stolfi for road design[15].

Given a set of points in feature space, we represent the set with a *signature*. The signature is composed of numbers of clusters of similar features in a Euclidean space. Each cluster is attached to a weight, which reflects the ratio of the number of features in this cluster to the total number of features in the set. During the process of video-based face recognition, each video corresponds to a feature distribution in feature space and it can be modelled as a signature. For simplicity and efficiency, we apply K-Means algorithm[14] for clustering. Each cluster contributes a pair $(\mu, p_\mu)$, where $\mu$ is the mean of the cluster and $p_\mu$ is the weight of the cluster. For videos, poses and expressions change constantly. The images in a video form a complex manifold 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 feature distribution in a set. 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.

Assume two distributions $g(A)$ and $g(B)$. We can imagine $g(A)$ is a mass of earth, and $g(B)$ 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". Figure 1 shows an example with three piles of earth and two holes. When $g(A)$ and $g(B)$ are represented with signatures, EMD is defined

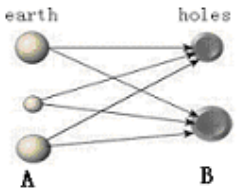


**Fig. 1.** An example of EMD

as the minimal "cost" needed to transform one signature to the other. EMD can be formalized as the following linear programming problem: Let $g(A) = \{(\mu_1, p_{\mu_1}), \ldots, (\mu_m, p_{\mu_m}\}$ and $g(B) = \{(\nu_1, p_{\nu_1}), \ldots, (\nu_n, p_{\nu_n}\}$, where $\mu_i$, $\nu_j$ are the mean vectors of clusters of $g(A)$ and $g(B)$, respectively, and $p_{\mu_i}$, $p_{\nu_j}$ is their corresponding weight. The cost to move an element $\mu_i$, to a new position $\nu_j$ 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 Euclidean distance between $\mu_i$ and $\nu_j$. EMD is the sum of cost of moving the weights of the elements of $g(A)$ to those of $g(B)$. Thus the solution to EMD is to find a set of cost coefficients $c_{ij}$ to minimize the following function:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}d_{ij} \tag{6}$$

subject to: (i) $c_{ij} \geq 0$, (ii) $\sum_{i=1}^{m} c_{ij} \leq p_{\nu_j}$, (iii) $\sum_{j=1}^{n} c_{ij} \leq p_{\mu_i}$, and (iv) $\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} = \min(\sum_{i=1}^{m} p_{\mu_i}, \sum_{j=1}^{n} p_{\nu_j})$. 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 its 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:

$$EMD(A,B) = EMD(g(A), g(B)) = \frac{\min(\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}d_{ij})}{\min(\sum_{i=1}^{m} p_{\mu_i}, \sum_{j=1}^{n} p_{\nu_j})} \tag{7}$$

As illuminated above, EMD reflects the average ground distance between two distributions. The cost of moving indicates the nearness of the signatures in Euclidian space. In our method, after linear transformation with FLDA, corresponding to each distribution, a signature is built with K-Mean algorithm as shown in Figure 2. Each signature contains a set of mean feature vectors and their corresponding weights. In Figure 2, the mean of each cluster is labelled with a red '$\star$' and the weight is denoted under the corresponding image. With more clusters are used, more precise the model is, and more difficult the problem of linear optimization is to solve. 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 cluster generally consist of faces under 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. Based on the above description, the similarity function between the training video $A$ and the testing video $B$ can be defined as:

$$f(A,B) = \exp(-\frac{EMD(A,B)}{\sigma^2}) \tag{8}$$

where $\sigma$ is a constant for normalization. The value of the function changes from 0 to 1. Bigger value means more similarity between $A$ and $B$.

## 3 Experimental Results

We take a combined database to evaluate the performance of our algorithm. Two experiments are performed. The first experiment fixes the sizes of image sets,

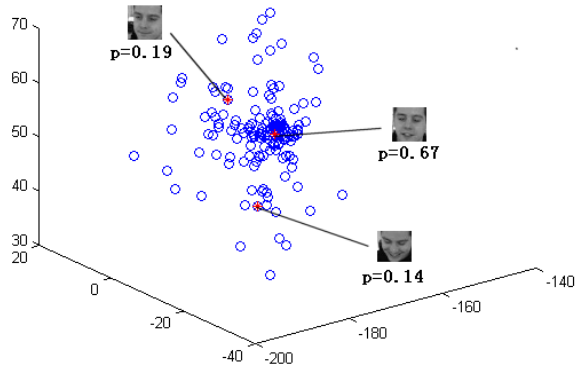


**Fig. 2.** A signature for video-based face recognition

and compares the recognition rate with changing the number of eigenvectors or features. The second experiment changes the sizes of image sets and records the recognition results. The methods are used for comparison is listed as follows:
- Mutual Subspace Method (MSM);
- K-L divergence for classification in original image space (K-L);
- K-L divergence in FLDA feature space (FLDA+K-L);
- EMD in FLDA feature space (FLDA+EMD);
- EMD in original image space (EMD).

We apply the following experimental methods for these algorithms. For MSM, K-L and EMD, we evaluate their performance directly in high dimensional image space. For the other two methods, we firstly reduce dimensionality based on FLDA. For FLDA+K-L, Gaussian function is used to model the set of feature data and K-L divergence between Gaussian models are estimated for classification. For FLDA+EMD, video's matching is based on the measurement of Earth Mover's Distance in the reduced dimensionality space. The label $K$ is assigned to the testing video if the following formula is satisfied:

$$K = argmax_A f(A, B) \tag{9}$$

where $A$ is the reference video in training sets, $B$ is the querying video, and $f(A, B)$ is the similarity function.

### 3.1 The Combined 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 25 subjects in the database. Totally, our combined database contains 50 subjects, and each subject has 300 face images. Figure 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.

**Fig. 3.** Some cropped faces from sequences

### 3.2 Recognition Rate vs. Number of Eigenvectors or Features

In this experiment, 60 frames of a video are for training and the remaining are for testing. The recognition results are shown in Figure 4. In Figure 4, the remaining frames in a video are divided into 4 testing sets. Each set contains 60 frames. The number of features is changing from 2 to 24. When more features are used, no changes can be observed.

Three methods, i.e., MSM, FLDA+K-L, FLDA+EMD, are compared in the experiment. Those methods have a common that the similarity function can be computed with changeable number of eigenvectors or features. For MSM, we change the number of eigenvectors to obtain the recognition rate. For other methods, we change that of features. When we use EMD for matching, only 3 clusters in a signature are used. Even more clusters are taken, no significant improvements of recognition rate are made. From Figure 4, we note that the recognition performance of FLDA+EMD is the best. Especially when less than 8 eigenvectors or features are for the experiment, FLDA+EMD outperforms MSM and FLDA+K-L. It is worth noting that the reason why FLDA+EMD is better than FLDA+K-L is that the Gaussian model for K-L divergence is too

simple to reflect the data distribution in feature space, while the signatures are competent for this task. We also note that MSM is better than FLDA+K-L for less than 4 or 5 eigenvectors or features. With the increasing of eigenvectors or features, FLDA+K-L will be better.

### 3.3 Recognition Rate vs. Size of Sets

In the experiment, fixing the number of eigenvectors as their maximal value and changing the size of the sets, we evaluate all the five algorithms in image space. They are: MSM, K-L, FLDA+K-L, FLDA+EMD and EMD. The different partition method of the sets in a video are listed as follows:
(i). A set of 60 images is for training, 8 sets of 30 images are for testing;
(ii). A set of 60 images is for training, 4 sets of 60 images are for testing;
(iii). A set of 100 images is for training, 5 sets of 40 images are for testing;
(iv). A set of 100 images is for training, 4 sets of 50 images are for testing.

The recognition result is shown in Table 1. From this table, we know that the recognition rate of FLDA+EMD is higher than the others. We also note that FLDA+K-L is better than K-L and FLDA+EMD is better than EMD. This phenomenon demonstrates that FLDA is an effective method to reduce dimension and make the features more discriminative. In addition, though EMD directly in image space is not comparable to MSM, but it is superior over K-L. It also demonstrates EMD is an effective metric for classification.

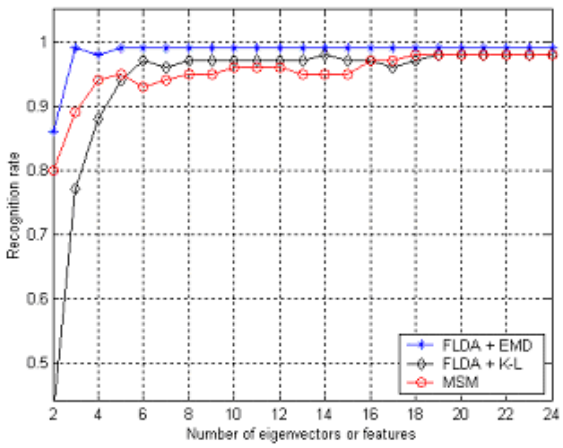


**Fig. 4.** Recognition rate vs. Number of features

**Table 1.** Recognition rate vs. Size of sets

| Training size | Testing size | MSM | K-L | FLDA+K-L | FLDA+EMD | EMD |
|---|---|---|---|---|---|---|
| 60 | 8×30 | 95% | 70.5% | 96% | 98% | 87% |
| 60 | 4×60 | 98% | 66% | 98% | 99% | 90% |
| 100 | 5×40 | 97% | 90% | 98% | 99% | 91% |
| 100 | 4×50 | 97% | 90% | 97% | 100% | 92% |

## 4 Conclusions

In this paper, we consider a most straightforward method of using distance for matching. The similarity function is established based on the computation of Earth Mover's Distance (EMD) between two distributions. The features are obtained by mapping the images from high dimensional image space to low dimensional FLDA feature space. Each set is represented with a signature. The solution to EMD is a linear optimization problem to find the minimal work needed to fill up one signature with the other. Experimental results show the performance of EMD and compare it to other batch methods. In future, we will consider the updating 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.

## Acknowledgements

This work is funded by research grants from the National Basic Research Program of China (No. 2004CB318110) and the National Natural Science Foundation of China (No. 60332010).

## 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] P.N.Belhumeur, J.P.Hespanha, D.J.Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997.

[8] 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.

[9] B. Moghaddam, A. Pentland, "Probabilistic visual learning for object representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, no.7, pp. 696-710, 1997.

[10] M.Turk, A.Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.

[11] M.S.Bartlett, H.M.Lades and T.Sejnowski, "Independent Component Representations for Face Recognition", In Proceedings of SPIE, 2399(3), pp. 528-539, 1998.

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

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

[14] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.

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

[16] 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.

[17] 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.