Video-based face recognition using manifold learning by Neural Networks

Kian Hamedani
Biomedical engeeniring department
Amirkabir university of technology
Tehran,Iran

Seyed Ali SeyedSalehi
Biomedical engeeniring department
Amirkabir university of technology
Tehran,Iran

***Abstract*— This paper proposes a novel method using manifold learning by neural networks for identifying people while they are talking. In training phase people say the same sentence,we train the NN for learning low-dimensional nonlinear manifolds that are embedded in high-dimensional video space.After training phase we use another video of the same persons while they are saying another sentence for testing.Comparing the recognition results with other methods shows that our method outperforms other methods.Finally we achieve 98.4% of recognition rate.**

Keywords— Neural Networks;manifold;face recognition ; video

I. INTRODUCTION

FOr decades human face recognition has been an active topic in the field of object recognition.A general statement of this problem can be formulated as follows:Given still or video images of a scene,identify person or persons in the scene[5].

Most of the algorithms have been proposed to deal with individual images ,where usually both the training and testing sets consist of facial pictures.The recognition performance of image based methods have been severly affected by different kinds of varations,like pose, illumination and expression changes.Thus researchers have started to look at video- based face recognition,in which both the enrolment and recognition sets are facial video sequences representing the client of the system[1].Currently ,most person recognition using videos are straightforward generalization of image-based algorithms; in these systems, the feature extraction and classification ,are applied independently to each frame,then the similarity scores are integrated using post-mapping information fusion techniques[1].

Video-based face recognition techniques are divided into two categories:those that neglect the temporal information and those that exploit it even partially[1].The most studied scenario in video-based face recognition is having a set of still images as gallery(training set) and video sequences as the probe(test set). Obviously, such an approach is not optimal as some important information in the video sequences may be left out.However,in some real-world applications such as in human-computer interaction and content based video retrieval, both training and test sets can be video sequences. In such settings, performing video-to-video matching may be crucial for robust face recognition but this task is far from being trivial[2].

The most important approaches neglecting the temporal information are eigenfaces(PCA), The eigenface technique is based on the notion of dimensionality reduction; in fact, Kirby and Sirovich were the first to remark that the dimensionality of the face space is much smaller of that of a single face, considered as an arbitrary image. A first method to reduce the image space into a low-dimensional feature space is to apply the principal component analysis[11]. In [12], Satoh proposed a straightforward extension of the traditional eigenface approach, by introducing a new similarity measure for matching video data. The similarity between distinct videos was obtained by considering the smallest distance between frame pairs (one from each video), in the reduced feature space. Fisherfaces(LDA) and LBP,In these methods no time and dynamic relation is considered between frames. Fisherfaces [13] is another state of the art technique for person recognition using facial appearance. Similarly to the eigenface approach fisherfaces is also based on the notion of face space reduction into a low-dimensional feature space. The optimal projection is calculated by applying the Fisher's linear discriminant (FLD) (also called linear discriminant analysis (LDA)).In [12], Satoh also proposed a straightforward extension of the traditional fisherface approach, by employing the same video similarity measure developed for the eigenface case. Again, the similarity between distinct videos was calculated by considering the smallest distance between frame pairs (one from each video), in the reduced feature space. HMM and ARMA are two methods which use dynamic information for video-based face recognition.Specially in these methods,HMM is the most famous one.In video-based face recognition using HMM,as shown in figure 1,during the

training process, the statistics of training video sequences of each subject, and the temporal dynamics, are learned by an HMM. During the recognition process, the temporal characteristics of the test video sequence are analyzed over time by the HMM corresponding to each subject. The likelihood scores provided by the HMMs are compared, and the highest score provides the identity of the test video sequence[3].

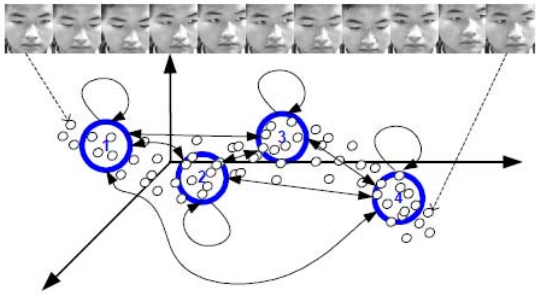

Fig. 1. Temporal HMM for modeling face sequences

In ARMA method we model a moving face as a linera-dynamical system whose appearance changes with pose. An autoregressive and moving average (ARMA) model is used to represent such a system. The choice of ARMA model is based on its ability to take care of the change in appearance while modeling the dynamics of pose, expression etc. Recognitionis performed using the concept of subspace angles to compute distances between probe and gallery video sequences[4].

In this article at first we introduce the database that we have used then will explain how to preprocess and prepare it for recognition.In this paper using the capabilities of neural networks in nonlinear signal processing ,some architecures of neural networks for analyzing nonlinear principle components have been propsed.At first a simple classifier multi-layer perceptron is used for face recognition of people who are speaking from video images then another architecture is propsed which can separate and learn low-dimensional nonlinear manifolds that are embeded in high dimensional video space ,this network learns both person and pose manifolds in its hidden supervised layer.In comparison with simple classifier,this network increase recognition rate about 4%.After that we train another network which unlike the last one learns pose manifold unsupervised,at this time again the recognition rate increases and 98.4% of recognition rate is acheived which is a high recognition rate. At the end we compare our method with other state of the art methods and it shows that our method outperforms them.

II. DATA INTRODUCTION AND PREPROCESSING

A. *Database introduction*

The database that we have used in this article is publicly available video database,VidTiMit.This database contains videos of 43 different persons.In a part of this database people say the same sentence and in a room their videos have been recorded(Fig2)[17].

Fig. 2. VidTiMit video database(talking people)

B. *Face detection*

Because in this database as is obvious in figure 3 face area is not completely cropped so we need a method which automaicly detects and crops the face area.A classification-based face detection method using Gabor filter features is proposed.Considering the desirable characteristics of spatial locality and orientation selectivities of the Gabor filter, we design filters for extracting facial features from the local image. The feature vector based on Gabor filters is used as the input of the classifier, which is a Feed Forward neural network (FFNN) on a reduced feature subspace.The image will be convolved with Gabor filters by multiplying the image by Gabor filters in frequency domain[7].

An image can be represented by the Gabor wavelet transform allowing the description of both the spatial frequency structure and spatial relations. Convolving the image with complex Gabor filters with 5 spatial frequency (v = 0,…,4) and 8 orientation (μ = 0,…,7) captures the whole frequency spectrum, both amplitude and phase.Ater this step it is time to extract the features. Feature vectors are extracted from points with high information content on the face image . From the responses of the face image to Gabor filters, peaks are found by searching the locations in a window *W0* of size *WxW* by the following procedure:

A feature point is located at *(x0, y0)*, if

$$R_j(x_0, y_0) = \max_{(x,y)\in W0}(R_j(x, y))$$

$$R_j(x_0, y_0) > \frac{1}{N1N2}\sum_{x=1}^{N1}\sum_{y=1}^{N2} R_j(x, y).$$

$$j = 1,2,\ldots\ldots\ldots,40$$

where *Rj* is the response of the face image to the *jth* Gabor filter *N1 N2* is the size of face peaks of the responses. In our experiment a *9x9* window is used to search feature points on Gabor filter responses. A feature map is constructed for the face by applying above process to each of 40 Gabor filters[7].

As it is seen in figure 3 the face area is completely cropped.This area is shown by green frame in figure 3.Since in this video database people only talk so by determining the face area in one frame the face area is determined in other frames.

Fig. 3.Face detection using gabor wavelet

After detecting the face area in all videos the resolution of videos becomes 60x60.

III. NEURAL NETWORK ARCHITECTURES

A. Simple classifier Neural Network

Here we use a simple multi-layer perceptron as a classifier and its structure is shown in figure 4.This network has 3600+1 inputs in first layer that equals to the number of the pixels of training images plus one node special for biasing.In the second and third layer there are respectively 400+1 and 200+1 neurons and finally in the fourth layer there are 43 neurons which is equal to the number of the persons that we use their videos.The back propagation algorithm is used for training this network.Learning coefficient is 0.001 and momentum factor is 0.7.

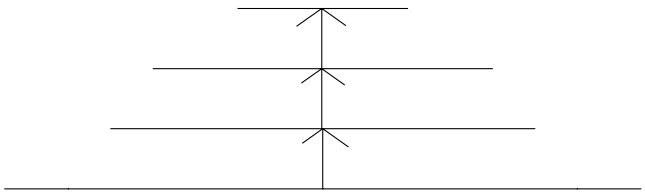

Fig. 4.classifier neural network architecture

In this part no dynamic relation is considered between the frames of videos.In the output layer there are 43 neurons,For every person a label is considered as the desired output.There are 43 bits in the output and for training the video of every person we make one bit 1 and let the other bits be zero for example the desired output for tarinig the video sample of the first person is 10000…..00000 and the desired output for the second person is 01000…..00000 and so on. After designing such architecture the network is trained by videos of 43 different persons in which they all say the same sentence.The error function of this training is shown in figure 5.

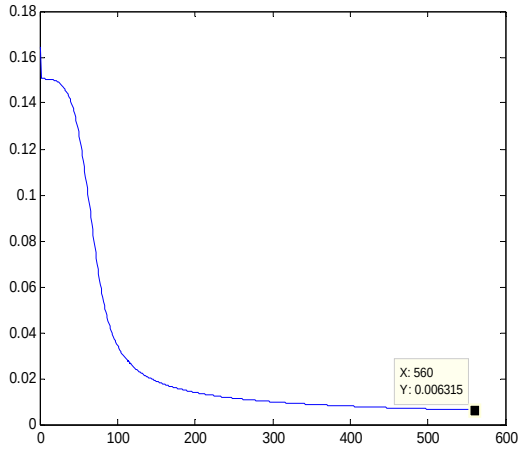

Fig. 5.error function for training the mentioned neural network

As it can be seen in the figure after about 560 itterations the error becomes about .006.The recognition rate for the training data was 100% this implies that the learning step has been done well.

Now for testing we use the videos of the same persons while they are saying another sentence.In this situation 91.12% of recognition rate is achieved,it means that 91.12% of all video frames have been recognized.9% reduction in recognition rate is caused because of the difference between the sentences in training and testing stage.As it can be seen in table 1 although the network that we have used in this part is simple it has achieved a high recognition rate which is comparable to other state of the art methods.In the next parts more complex architectures of neural networks are introduced that can sepearte and learn pose and person manifolds in their hidden layer and in this way the effect of pose changing that here is caused by speaking is nonlinearly filtered from the identity of people and so the recognition rate is increased.

B. Designing a network for learning and seperating pose and person manifolds

The face image contains two types of independent information "Person information"and "Pose information" which person information contains identity information.Here pose information contains the lip position of the speaker while speaking. $\vec{x}$ is the n dimensional vector of a person image in one pose,this image in m dimensional input space is shown by a dot.As this pose starts to change here it means that the person starts to talk,this dot starts to move in the input space and the manifold that here is produced is called the pose changing manifold of this person[15].

So for learning and seperating pose and person manifolds an autoassociative neural network can be used for analyzing nonlinear principle components that seperates pose features from person features.In figure 6 an architecture for this network has been shown.This network is composed of two parts,its first part is analyzing part which projects whole the videos into person and pose manifolds ,the second part is for synthesizing

input frames in the output,this is done by combining nonlinear manifolds.This network can approximately reconstruct input frames so it is an autoassociative neural network[15],[16].

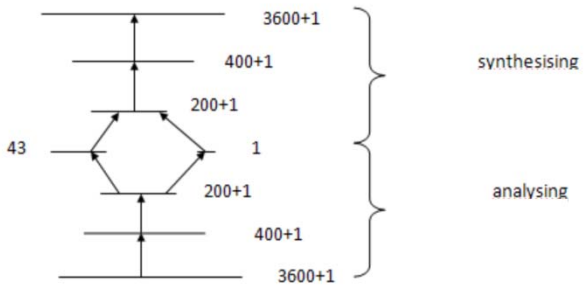

Fig. 6.pose and person separating neural network

The layer with 43 neurons learns person manifold and the layer with 1 neuron learns pose manifold,in proceeding we will describe how.Here both person and pose manifolds are learned supervised,it means that some labels are considered for these manifolds.The person codes are considered like the last network it means 100000….0000 for the first person and so on but for the pose manifold since the videos are not aligned together so an unique code cant be considered for all poses.But as it is obvious in figure 2 we suppose that during speaking a person's face is approximately in one pose and the face pose changes a little because while speaking face pose doesn't change severly and speaker opens and closes his or her lips alittle and the expression of the other parts of face doesn't change(figure 2),so in the pose layer one supervised neuron is trained to learn the common pose of the persons while speaking and separate it from person manifold.

As the error function of this network shows(figure7) at first the network refuses to learn the manifolds in such architecture and even in its first 100 itterations error starts to increase but since here pose changing is not so sever at last the neural network is forced to learn a unique pose as the pose manifold of the persons while speaking .In this way the pose and person manifolds are separated to some extent and the recognition rate is increased.Here one label for example 1 is considered for all poses.

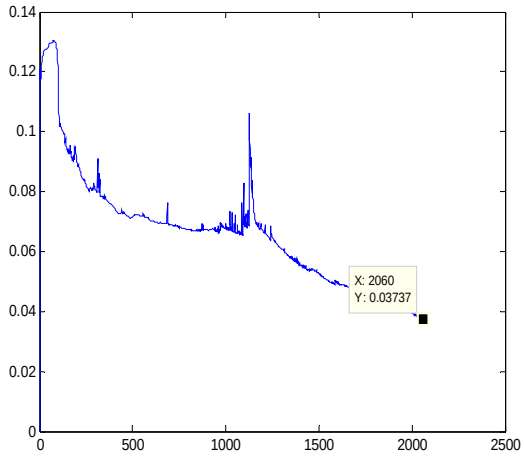

Fig. 7.error function for training the above neural network

The error of network after about 2060 itterations becomes about .037. The recognition rate for the training data was 100% . For testing we use the videos of the same persons while they are saying another sentence.In this situation 95.47% of recognition rate is achieved which is 4% more than the recognition rate in the previous network,this increment in recognition rate shows the capability of this network in separating pose and person manifolds.After this we reconstruct some of the synthesized videos,the result is shown in figure 8,as it can be seen the synthesized output is just a single frame which is the normal image of the speaker.

Fig. 8.Synthesized output for an input video sample

C. Unsupervised training of pose manifold

In the last part one code is considered as the pose label and the network is forced to project all poses to one pose,this bound imposes a lot of pressure to network so in this part we decided to consider no label for the pose and let the neural network itself form the pose manifold,but here again the person manifold is trained supervised like two last parts.To achieve this purpose the network's architecture is changed in hidden layer in this way,43 neurons are trained supervised for person manifold and 5 unsupervised neurons for pose manifold. Here the network is free to form the pose manifold.NN has to minimize the output error by adjusting the weights of layers and at the end generates proper codes for poses in the unsupervised hidden layer and reconstructs video frames in the

output. As it can be seen in the error function of training this network(fig9) ,the error of this network converges much easier than the previous network.

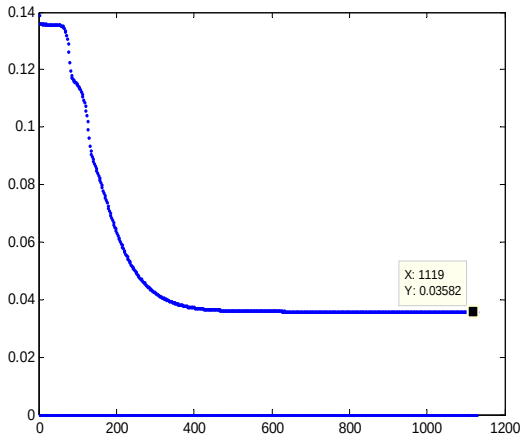

Fig. 9.error function for training unsupervised pose network

After training and gaining proper weights for each layer , we test this method and achieve 98.4% of recognition rate which is 3% more than the last network.As it is shown in table1 this method outperforms many state of the art methods.In order to show this increment in another way some synthesized videos in the output are reconstructed,in the reconstructed videos unlike the last network people's lips move and the reconstructed videos are not just single frames and again in this way our assumption is proved and better reconstruction of videos in the output shows that pose manifold in this architecture is better learnt and separated from person manifold.

IV. COMPARISON WITH OTHER METHODS

In table 1 the comparison between our method and other state of the art methods is written[2].As you can see our method achieves higher recognition rate than the other methods,the interesting point in this evaluation is that some methods like HMM and ARMA although they use dynamic information but they don't achieve higher recognition rates,maybe it is because these methods only use global information and don't pay attention to local information[2].

The drawback of PCA method is that taking into account the large variation of facial appearance in a given sequence, the choice of the closest frame pair showed limited robustness to outliers and obtained poor results[12]. The LDA strategy experienced the same weaknesses to outlier frames as before, but it obtained better recognition results because the fisherface method is known to be more discriminating than the eigenface one[13].

Because video sequences are short HMM gave poor results. This is probably due to the fact that facial movement is learned more slowly than the static facial structure. Importantly, one may not expect worse results using the spatio-temporal representations. However, the obtained results attest that PCA and LDA-based representations perform better in such cases. This means that the spatio-temporal representation did not succeed to discover the importance of the spatial cue over its temporal counterpart[14].

TABLE I.

method	**Recognition rate%**	**resolution**
PCA	93.2%	60x60
LDA	94%	60x60
LBP	97.6%	60x60
HMM	92.9%	60x60
ARMA	95.8%	60x60
simple Classifier NN	91.12%	60x60
Supervised training of pose	95.47%	60x60
Unsupervised training of pose	98.4%	60x60

V. CONCLUSION

It is concluded that neural networks are good tools for video based face recognition,specially by designing architectures which can separate pose and person features in their hidden layer the recognition rate can be increased significantly,another benefit of separating pose and person manifolds is that in this way we can synthesize the virtual video of a single image. Another conclusion is that using spatio-temporal methods like HMM does not always gurantee better performance and their performance depends on the length of videos.

REFERENCES

[1] Fedrico, M., Dugely, J., *" Person recognition using facial video information: A state of the art "*, Journal of Visual Languages and Computing 20 (2009) 180–187.

[2] Hadid, A., Pietik¨ainen, M., *"Manifold learning for video-video face recognition "*, Springer-Verlag Berlin Heidelberg 2009.

[3] Liu, X., Chen, T., *"Video-based face recognition using adaptive hidden markov models "*, IEEE Int. Conf. on Computer Vision and Pattern Recognition, June 2003, pp. 340–345 (2003) .

[4] Aggarwal, G., Chowdhury, A.R., Chellappa, R., *"A system identification approach for video-based face recognition*",In: 17th International Conference on Pattern Recognition, August 2004, vol. 4, pp. 175–178 (2004) .

[5] W. Zhao, R. Chellappa, P.J. Phillips and A. Rosenfeld," *Face recognition:a literature survey*", ACM Comput. Surv., December Issue, 2003. pp. 399-458.

[6] Seung, H.S., Lee, D, *"* The manifold ways of perception *"*Science 290(12), 2268–2269 (2000)

[7] Gupta, B., Gupta, S.,Tiwari.,A., *"Fcae detection using gabor feature extraction and artificial Neural Networks ",*

[8] Hadid, A., Pietik¨ainen, M., *" Manifold learning for gender classification from face sequences "*, In: Proc. 3rd IAPR/IEEE International Conference on Biometrics, ICB 2009 (2009).

[9] Wang, H., Wang, Y., Cao, Y., *"Video based face recognition:A survey"*, World Academy of Science, Engineering and Technology 60 2009

[10] S. Lawrence, C.L. Giles, A. Tsoi and A. Back, *"Face recognition: a convolutional neural-network approach"*, IEEE Trans. Neural Networks, 8 (1), 1997, pp. 98-113.

[11] M.A. Turk, A.P. Pentland, *"*Face recognition using eigenfaces*"*, in: IEEE Proceedings on Computer Vision and Pattern Recognition, 1991, 586–591.

[12] S. Satoh, *"*Comparative evaluation of face sequence matching for content-based video*"* access, in: IEEE Proceedings on Automatic Face and Gesture Recognition, 2000, pp. 163–168.

[13] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, *"*Eigenfaces vs. fish-erfaces: recognition using class specific linear projection*"*, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[14] Hadid, A., Pietik¨ainen, M , *"*From Still Image to Video-Based Face Recognition: An Experimental Analysis*"*, Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition 2004.

[15] Dadashi,N, *"*Face recognition from single image per person*"*,A thesis submitted to the master of science degree ,the department of biomedical engineering,Amirkabir university of technology,December 2008.

[16] Huang, F.,Zhou, Z.,Zhang,H.,Chen, T, *"*Pose varied face recognition*"*,IEEE Intl. Conf. On Automatic Face and Gesture Recognition, Grenoble, France, 2000

[17] http:// VidTiMit database from: http://www.itee.uq.edu.au/~conrad/vidtimit/