Sparse Representation for Video-Based Face Recognition

Imran Naseem[1], Roberto Togneri[1], and Mohammed Bennamoun[2]

[1] School of Electrical, Electronic and Computer Engineering
The University of Western Australia
imran.naseem@ee.uwa.edu.au, roberto@ee.uwa.edu.au
[2] School of Computer Science and Software Engineering
The University of Western Australia
bennamou@csse.uwa.edu.au

Abstract. In this paper we address for the first time, the problem of video-based face recognition in the context of sparse representation classification (SRC). The SRC classification using still face images, has recently emerged as a new paradigm in the research of view-based face recognition. In this research we extend the SRC algorithm for the problem of temporal face recognition. Extensive identification and verification experiments were conducted using the VidTIMIT database [1,2]. Comparative analysis with state-of-the-art Scale Invariant Feature Transform (SIFT) based recognition was also performed. The SRC algorithm achieved 94.45% recognition accuracy which was found comparable to 93.83% results for the SIFT based approach. Verification experiments yielded 1.30% Equal Error Rate (EER) for the SRC which outperformed the SIFT approach by a margin of 0.5%. Finally the two classifiers were fused using the weighted sum rule. The fusion results consistently outperformed the individual experts for identification, verification and rank-profile evaluation protocols.

1 Introduction

It is long known that appearance-based face recognition systems critically depend on manifold learning methods. A gray-scale face image of order $a \times b$ can be represented as an ab-dimensional vector in the original *image space*. However any attempt of recognition in such a high dimensional space is vulnerable to a variety of issues often referred to as the *curse of dimensionality*. Typically in pattern recognition problems it is believed that high-dimensional data vectors are redundant measurements of an underlying source. The objective of manifold learning is therefore to uncover this "underlying source" by a suitable transformation of high-dimensional measurements to low-dimensional data vectors. View-based face recognition methods are no exception to this rule. Therefore, at the feature extraction stage, images are transformed to low dimensional vectors in a *face space*. The main objective is to find a basis function for this transformation,

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 219–228, 2009.

which could distinguishably represent faces in the face space. Linear transformation from the image space to the feature space is perhaps the most traditional way of dimensionality-reduction, also called "Linear Subspace Analysis".

A number of approaches have been reported in the literature including Principal Component Analysis (PCA) [3], [4], Linear Discriminant Analysis (LDA) [5] and Independent Component Analysis (ICA) [6], [7]. These approaches have been classified in two categories namely *reconstructive* and *discriminative* methods. Reconstructive approaches (such as PCA and ICA) are reported to be robust for the problem related to contaminated pixels, whereas discriminative approaches (such as LDA) are known to yield better results in clean conditions [8]. Nevertheless, the choice of the manifold learning method for a given problem of face recognition has been a hot topic of research in the face recognition literature. These debates have recently been challenged by a new concept of "Sparse Representation Classification (SRC)" [9]. It has been shown that unorthodox features such as downsampled images and random projections can serve equally well. As a result the choice of the feature space may no longer be so critical [9]. What really matters is the dimensionality of the feature space and the design of the classifier. The key factor to the success of sparse representation classification is the recent development of "Compressive Sensing" theory [10].

Due to the ever increasing security threats, video surveillance systems have been deployed on a large scale. With the additional temporal dimension, video sequences are much more informative than still images. As a result the person identification task is facilitated due to specific attributes of each subject such as head rotation and pose variation along the temporal dimension. Additionally more efficient face representations such as super resolution images can be derived from video sequences for further enhancement of the overall system. These motivations have urged researchers to look into the development of face recognition systems that can utilize the spatiotemporal information in video sequences. It is therefore becoming imperative to evaluate present state-of-the-art face recognition algorithms for video-based applications. With this understanding, this research is targeted to the extension of the recently proposed SRC classification for the problem of video-based face recognition. The primary objective is to critically analyze the new approach in comparison with state-of-the-art SIFT features based algorithm. The rest of the paper is organized as follows: Section 2 provides an overview of the SRC algorithm followed by a brief description of SIFT based recognition in Section 3. Experimental results and discussion are presented in Section 4, the paper concludes in Section 5.

2 Sparse Representation for Face Recognition

We now discuss the basic framework of the face recognition system in the context of sparse representation [9]. Let us assume that we have k distinct classes and n_i images available for training from the ith class. Each training sample is a gray scale image of order $a \times b$. The image is downsampled to an order $w \times h$

and is converted into a 1-D vector $\mathbf{v}_{i,j}$ by concatenating the columns of the downsampled image such that $\mathbf{v}_{i,j} \in \mathbb{R}^m$ $(m = wh)$. Here i is the index of the class, $i = 1, 2, \ldots, k$ and j is the index of the training sample, $j = 1, 2, \ldots, n_i$. All this training data from the ith class is placed in a matrix A_i such that $A_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots\ldots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$. As stated in [9], when the training samples from the ith class are sufficient, the test sample $\mathbf{y}$ from the same class will approximately lie in the linear span of the columns of A_i:

$$\mathbf{y} = \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + \cdots + \alpha_{i,n_i}\mathbf{v}_{i,n_i} \tag{1}$$

where $\alpha_{i,j}$ are real scalar quantities. Now we develop a dictionary matrix A for all k classes by concatenating A_i, $i = 1, 2, \ldots, k$ as follows:

$$\mathbf{A} = [A_1, A_2, \ldots, A_k] \ \in \mathbb{R}^{m \times n_i k} \tag{2}$$

Now a test pattern $\mathbf{y}$ can be represented as a linear combination of all n training samples $(n = n_i \times k)$:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{3}$$

Where $\mathbf{x}$ is an unknown vector of coefficients. Now from equation 3 it is relatively straight forward to note that only those entries of $\mathbf{x}$ that are non-zero correspond to the class of $\mathbf{y}$ [9]. This means that if we are able to solve equation 3 for $\mathbf{x}$ we can actually find the class of the test pattern $\mathbf{y}$. Recent research in compressive sensing and sparse representation [11,10,12,13,14] have shown that using the sparsity of the solution of equation 3, enables us to solve the problem using l_1-norm minimization:

$$(l^1): \quad \hat{\mathbf{x}}_1 = argmin \left\|\mathbf{x}\right\|_1 \quad ; \mathbf{A}\mathbf{x} = \mathbf{y} \tag{4}$$

Once we have estimated $\hat{\mathbf{x}}_1$, ideally it should have nonzero entries corresponding to the class of $\mathbf{y}$ and now deciding the class of $\mathbf{y}$ is a simple matter of locating indices of the non-zero entries in $\hat{\mathbf{x}}_1$. However due to noise and modeling limitations $\hat{\mathbf{x}}_1$ is commonly corrupted by some small nonzero entries belonging to different classes. To resolve this problem we define an operator δ_i for each class i so that $\delta_i(\hat{\mathbf{x}}_1)$ gives us a vector $\in \mathbb{R}^n$ where the only nonzero entries are from the ith class. This process is repeated k times for each class. Now for a given class i we can approximate $\hat{\mathbf{y}}_i = \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)$ and assign the test pattern to the class with a minimum residual between $\mathbf{y}$ and $\hat{\mathbf{y}}_i$.

$$\underbrace{\min}_{i} r_i(\mathbf{y}) = \left\|\mathbf{y} - \mathbf{A}\delta_i(\hat{\mathbf{x}}_1)\right\|_2 \tag{5}$$

3 Scale Invariant Feature Transform (SIFT) for Face Recognition

The Scale Invariant Feature Transform (SIFT) was proposed in 1999 for the extraction of unique features from images [15]. The idea, initially proposed for

a more generic object recognition task, was later successfully applied for the problem of face recognition [16]. Interesting characteristics of scale/rotation invariance and locality in both spatial and frequency domains have made the SIFT-based approach a pretty much standard technique in the paradigm of view-based face recognition. The first step in the derivation of the SIFT features is the identification of potential pixels of interest called "keypoints", in the face image. An efficient away of achieving this is to make use of the scale-space extrema of the Difference-of-Gaussian (DoG) function convolved with the face image [15]. These potential keypoints are further refined based on the high contrasts, good localization along edges and the ratio of principal curvatures criterion. Orientation(s) are then assigned to each keypoint based on local image gradient direction(s). A gradient orientation histogram is formed using the neighboring pixels of each keypoint. Contribution from neighbors are weighted by their magnitudes and by a circular Gaussian window. Peaks in the histogram represent the dominant directions and are used to align the histogram for rotation invariance. 4×4 pixel neighborhoods are used to extract eight bin histograms resulting in 128-dimensional SIFT features. For illumination robustness, the vectors are normalized to unity, thresholded to a ceiling of 0.2 and finally renormalized to unit length. Figure 1 shows a typical face from the VidTIMIT database [1,2] with extracted SIFT features.

Fig. 1. A typical localized face from the VidTIMIT database with extracted SIFTs

During validation a SIFT feature vector from the query video f_q is matched with the feature vector from the gallery:

$$e = \arccos\left[f_q(f_g)^T\right] \tag{6}$$

where f_g corresponds to a SIFT vector from a training video sequence. All SIFT vectors from the query frame are matched with all SIFT features from a training frame using Equation 6. Pairs of features with the minimum error e are considered as matches. Note that if more than one SIFT vector from a given query

frame happens to be the best match with the same SIFT vector from gallery (i.e. many-to-one match scenario), the one with the minimum error e is chosen. Other false matches were reduced by matching the SIFT vectors from only nearby regions of the two images.

In principle, for different image pairs we have different number of matches. This information is further harnessed to be used as an additional similarity measure between the two faces. The final similarity score between two frames is computed by normalizing the average error $\overline{e}$ between their matching pairs of SIFT features and the total number of matches z on a scale [0,1] and then using a weighted sum rule.

$$\overline{e}' = \frac{\overline{e} - \min(\overline{e})}{\max(\overline{e} - \min(\overline{e}))} \tag{7}$$

$$z' = \frac{z - \min(z)}{\max(z - \min(z))} \tag{8}$$

$$s = \frac{1}{2}(\beta_e \overline{e}' + \beta_z(1 - z')) \tag{9}$$

where β_e and β_z are the weights of normalized average error $\overline{e}'$ and normalized number of matches z' respectively. It has to be noted that $\overline{e}'$ is a distance (dissimilarity) measure while z' is a similarity score, therefore in Equation 9 z' is subtracted from 1 for a homogeneous fusion. Consequently s becomes a distance measure.

4 Results and Discussion

The problem of temporal face recognition using the SRC and SIFT feature based face recognition algorithms was evaluated on the VidTIMIT database [1], [2]. VidTIMIT is a multimodal database consisting of video sequences and corresponding audio files from 43 distinct subjects. The video section of the database characterizes 10 different video files from each subject. Each video file is a sequence of 512×384 JPEG images. Two video sequences were used for training while the remaining eight were used for validation. Due to the high correlation between consecutive frames, training and testing were carried out on alternate frames. Off-line batch learning mode [17] was used for these experiments and therefore probe frames did not add any information to the system.

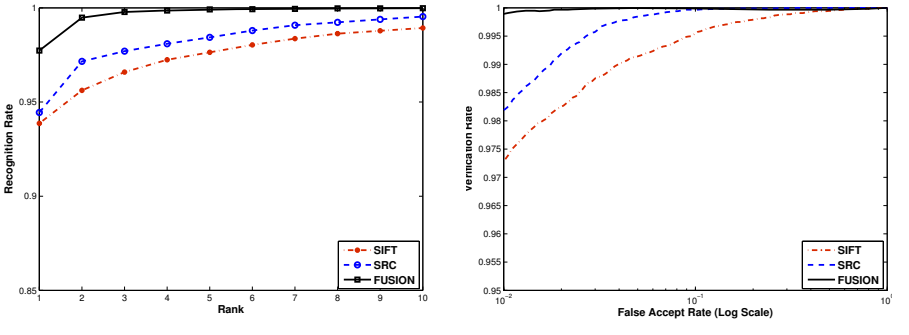
Face localization is the first step in any face recognition system. Fully automatic face localization was carried out using a Harr-like feature based face detection algorithm [18] during off-line training and on-line recognition sessions. For the SIFT based face recognition, each detected face in a video frame was scale-normalized to 150×150 and histogram equalized before the extraction of the SIFT features. We achieved an identification rate of 93.83%. Verification experiments were also conducted for a more comprehensive comparison between

Fig. 2. A sample video sequence from the VidTIMIT database

the two approaches. An Equal Error Rate (EER) of 1.8% was achieved for the SIFT based verification. Verification rate at 0.01 False Accept Rate (FAR) was found to be 97.32%.

For the SRC classifier, each detected face in a frame is downsampled to order 10×10. Column concatenation is carried out to generate a 100-dimensional feature vector as discussed in Section 2. Off-line batch learning is carried out on alternate frames using two video sequences as discussed above. Unorthodox downsampled images in combination with the SRC classifier yielded quite comparable recognition accuracy of 94.45%. EER dropped to 1.3% with a verification accuracy of 98.23% at 0.01 FAR. The rank profile and ROC (Receiver Operating Characteristics) curves are shown in Figure 3 (a) and 3 (b) respectively.

We further investigated the complementary nature of the two classifiers by fusing them at the score level. The weighted sum rule is used which is perhaps



(a) (b)

Fig. 3. (a)Rank profiles and (b) ROC curves for the SIFT, SRC and the combination of the two classifiers

the major work-horse in the field of combining classifiers [19]. Both classifiers were equally weighted and a high recognition accuracy of 97.73% was achieved which outperforms the SIFT based classifier and the SRC classifier by a margin of 3.90% and 3.28% respectively. Verification experiments also produced superior results with an EER of 0.3% which is better than the SIFT and the SRC based classification by 1.5% and 1.0% respectively. An excellent verification of 99.90% at an FAR of 0.01 is reported. Fusion of the two classifiers substantially improved the rank profile as well achieving 100% results at rank-5 only. A detailed comparison of the results is provided in Table 1.

Table 1. Summary of results

Evaluation Attributes	**SIFT**	**SRC**	**Fusion**
Recognition Accuracy	93.83%	94.45%	97.73%
Equal Error Rate	1.80%	1.30%	0.30%
Verification rate at 0.01 FAR	97.32%	98.23%	99.90%

Presented results certainly reflect a comparable performance index for the SRC classifier as compared to state-of-the-art SIFT based recognition. Extensive experiments based on identification, verification and rank-recognition evaluations consistently reflect better results for the SRC approach. Moreover the complementary information exhibited by the SRC method increased the verification success of the combined system to 99.9% for the standard 0.01 FAR criterion. Figure 4 shows variation in the recognition accuracy with the change in the normalized weight of the SRC classifier at the fusion stage. Approximately the highest recognition is achieved when both classifiers were equally weighted i.e. no prior information of the participating experts was incorporated in fusion.

Apart from these appreciable results it was found that the l_1-norm minimization using a large dictionary matrix made the iterative convergence lengthy and slow. To provide a comparative value we performed computational analysis for a randomly selected identification trail. The time required by the SRC algorithm for classifying a single frame on a typical 2.66 GHz machine with 2 GB memory was found to be 297.46 seconds (approximately 5 minutes). This duration is approximately 5 times greater than the processing time of the SIFT algorithm for the same frame which was found to be 58.18 seconds (approximately 1 minute). Typically a video sequence consists of hundreds of frames which would suggest a rather prolonged span for the evaluation of the whole video sequence. Noteworthy is the fact that experiments were conducted using an offline learning mode [17]. The probe frames did not contribute to the dictionary information. Critically speaking, the spatiotemporal information in video sequences is best harnessed using smart online [20] and hybrid [21] learning modes. These interactive learning algorithms add useful information along the temporal dimension and therefore enhance the overall performance. However, in the context of SRC classification, this would suggest an even larger dictionary matrix and consequently a lengthier evaluation.

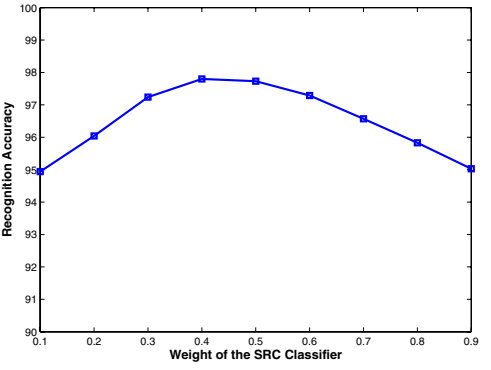


Fig. 4. Variation in performance with respect to bias in fusion

5 Conclusion

Sparse representation classification has recently emerged as the latest paradigm in the research of appearance-based face recognition. In this research we evaluated the approach for the problem of video-based face recognition. An identification rate of 94.45% is achieved on the VidTIMIT database which is quite comparable to 93.83% accuracy using state-of-the-art SIFT features based algorithm. Verification experiments were also conducted and the SRC approach exhibited an EER of 1.30% which is 0.5% better than the SIFT method. The SRC classifier was found to nicely complement the SIFT based method, the fusion of the two methods using the weighted sum rule consistently produced superior results for identification, verification and rank-recognition experiments. However since SRC requires an iterative convergence using an l_1-norm minimization, the approach was found computationally expensive as compared to the SIFT based recognition. Typically SRC required 5 minutes (approximately) for processing a single recognition trial which is 5 times greater than the time required by the SIFT based approach. To the best of our knowledge, this is the first evaluation of the SRC algorithm on a video database. From the experiments presented in the paper, it is quite safe to maintain that additional work is required before the SRC approach is declared as a standard approach for video-based applications. Computational expense is arguably an inherent issue with video processing giving rise to the emerging area of "Video Abstraction". Efficient algorithms have been proposed to cluster video sequences along the temporal dimension (for example [22] including others). These clusters are then portrayed by cluster-representative frame(s)/features resulting in a substantial decrease of complexity. Given the good performance of the SRC algorithm presented in this research, the evaluation of the method using state-of-the-art video abstraction methods will be the subject of our future research.

Acknowledgement

The authors would like to thank D. Lowe for providing the SIFT code. This research is partially funded by the Australian Research Council (ARC) grant No. DP0771294.

References

1. Sanderson, C., Paliwal, K.K.: Identity verification using speech and face information. Digital Signal Processing 14(5), 449–480 (2004)
2. Sanderson, C.: Biometric person recognition: Face, speech and fusion. VDM-Verlag (2008)
3. Jolliffe, I.T.: Pricipal Component Analysis. Springer, New York (1986)
4. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neurosicence 3(1), 71–86 (1991)
5. Belhumeur, V., Hespanha, J., Kriegman, D.: Eigenfaces vs Fisherfaces: Recognition using class specific linear projection. IEEE Tran. PAMI 17(7), 711–720 (1997)
6. Comon, P.: Independent Component Analysis - A New Concept? Signal Processing 36, 287–314 (1994)
7. Bartlett, M., Lades, H., Sejnowski, T.: Independent component representations for face recognition. In: Proceedings of the SPIE: Conference on Human Vision and Electronic Imaging III, vol. 3299, pp. 528–539 (1998)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, Inc., Chichester (2000)
9. Wright, J., Yang, A., Ganesh, A., Sastri, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. PAMI (2008)
10. Donoho, D.: Compressed sensing. IEEE Trans. Inform. Theory 52(4), 1289–1306 (2006)
11. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory 52(2), 489–509 (2006)
12. Donoho, D.: For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. Comm. on Pure and Applied Math. 59(6), 797–829 (2006)
13. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Comm. on Pure and Applied Math. 59(8), 1207–1223 (2006)
14. Candès, E., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Tran. Infm. Theory 52(12), 5406–5425 (2006)
15. Lowe, D.: Object recognition from local scale-invariant features. In: Intl. Conf. on Computer Vision, pp. 1150–1157 (1999)
16. Bicego, M., Lagorio, A., Grosso, E., Tistarelli, M.: On the use of SIFT features for face authentication. In: CVPRW (2006)
17. Lee, K., Ho, J., Yang, M., Kriegman, D.: Visual tracking and recognition using probabilistic appearance manifolds. CVIU 99(3), 303–331 (2005)
18. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)

19. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(3), 226–238 (1998)
20. Liu, L., Wang, Y., Tan, T.: Online appearance model. In: CVPR, pp. 1–7 (2007)
21. Lee, K., Kriegman, D.: Online probabilistic appearance manifolds for video-based recognition and tracking. In: CVPR, vol. 1, pp. 852–859 (2005)
22. Chan, A.B., Vasconcelos, N.: Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Trans. PAMI 30, 909–926 (2008)