A NEAR OPTIMAL PROJECTION FOR SPARSE REPRESENTATION BASED CLASSIFICATION

Sreekanth Raja *R. Venkatesh Babu*

Video Analytics Lab
SERC, Indian Institute of Science
Bangalore, India

ABSTRACT

Sparse representation based classification (SRC) is one of the most successful methods that has been developed in recent times for face recognition. Optimal projection for Sparse representation based classification (OPSRC)[1] provides a dimensionality reduction map that is supposed to give optimum performance for SRC framework. However, the computational complexity involved in this method is too high. Here, we propose a new projection technique using the data scatter matrix which is computationally superior to the optimal projection method with comparable classification accuracy with respect OPSRC. The performance of the proposed approach is benchmarked with various publicly available face database.

Index Terms— Sparse Representation, Subspace Projection, Face Recognition

1. INTRODUCTION

Face recognition is one of the most inevitable parts of modern day biometric identification systems, along with fingerprint identification, iris based recognition etc. It has got wide spread applications in various military as well as civilian applications. There is a vast plethora of research literature available on face recognition techniques [2][3].

With the advent of compressed sensing theory [4][5], sparse representation is being successfully used for face recognition. Wright et al. [6] introduced the concept of sparse representation based classification (SRC) where, the test sample is represented as a sparse linear combination of the training images. The sparsity structure of the coefficients encodes the information about the identity of the test vector. However, there is slight difference between compressed sensing and the sparse representation based classification suggested by Wright et al. In compressed sensing theory, the main aim is to recover a signal completely by sampling at a sub Nyquist rate. On the other hand, in sparse representation based classification, the sparsity structure of the signal representation is used to decode the identity of an unknown signal.

Before the introduction of SRC, the most popular face recognition algorithms were based on subspace methods. Eigenfaces [7], Fisherfaces [8] and Laplacianfaces [9] were the most popular among them. These are dimension reduction techniques, which project the high dimension face data into a lower dimension $face\ subspace$. Final classification is done in this subspace. Random Projection [10] [11] combines the idea of SRC and subspace methods by projecting each face image into a random subspace. Lu [1] came up with a supervised dimension reduction algorithm that gives a projection that is supposed to be optimum for sparse representation based classification framework. However, this method is high on its computational complexity. In this paper, an attempt is made to club subspace method and SRC by developing a projection map that maps the high dimensional face space to a lower dimension face subspace, without compromising on the discriminatory nature of the data. The proposed discriminative projection for SRC is low on computation, and high on classification accuracy. Experimental results show that the proposed projection provides comparable performance as the of OP-SRC.

Section 2, introduces the sparse representation based classification framework. Section 3, presents various subspace projections for SRC including random projection, OP-SRC and the proposed projection. The experimental results are presented in section 4 and conclusion in section 5.

2. THE SPARSE REPRESENTATION FRAMEWORK

In face recognition problems, each face is treated as an $m \times n$ matrix, reshaped into an $mn \times 1$ vector. Assume there are k distinct classes of face data. Let $T_i = [t_{i,1}, t_{i,2}, \cdots t_{i,l_i}]$ be the collection of vectors that represent the i^{th} class. Assume, that there are sufficient number of training vectors for all the classes. Given any new arbitrary sample vector y of the i^{th} class, it can be approximated by a linear combination of the training vectors.

$$y = \sum_{j=1}^{l_i} a_{i,j} t_{i,j} \tag{1}$$

where, $a_{i,j}$ represents the weight (coefficient) of basis training vector $t_{i,j}$.

Now the problem in face recognition is to find the class i to which the test vector y actually belongs. For this we consider the concatenated dictionary matrix T

$$T = [T_1\ T_2 \ldots T_k] \tag{2}$$

The columns of the matrix T forms the dictionary bases. Now y can be written as

$$y = Tx \tag{3}$$

where, $x = [0, 0, ...a_{i,1}, a_{i,2}, ..a_{i,l_i}, 0, 0....0]$

The solution vector x is expected to encode the identity of the test vector y. Unlike the Nearest Neighbor (NN) classifier or the Nearest Subspace (NS) [12] classifier, SRC uses the entire training set at a time to solve for x. The components of x are zeros except for those associated with the i^{th} class. Now the entire problem reduces to the most fundamental problem of linear algebra - that of solving the system of equation $Tx = y$. In practice, (3) is

an under-determined system, since the total number of training vectors is much more than the size of the vector. In order to avoid the anomaly of inconsistency of the system, we assume that the matrix T has full rank. Thus the system (3) gives an infinite number of solutions. The conventional l_2 solution for the problem is given by:

$$\widehat{x} = \arg\min \|x\|_2 \text{ subject to } Tx = y \tag{4}$$

This system can easily be solved using the pseudo inverse of T. However the solution can be *dense* i.e, there can be a large number of non-zero entries corresponding to coefficients of other classes and hence, may not be of much use in getting the identity of y. Hence l_2 solution is not suitable for this kind of problem. Since the test vector is represented using the training vectors from the same class only, we are looking for a *sparse* solution, i.e., a solution with minimal l_0 norm. Though l_0 norm do not follow the strict definition of a norm, it is defined as the number of nonzero entries in a vector. The identity of y is determined by the sparsity structure of x. Thus the problem is redefined as:

$$\widehat{x} = \arg\min \|x\|_0 \text{ subject to } Tx = y \tag{5}$$

Theoretically, if the sparsity of the solution is less than $mn/2$, this is the most optimum sparse solution which one can obtain [13]. But this is an NP hard problem. However if the solution is sufficiently sparse, the solution is equal to that of the following l_1 minimization problem that can be solved in polynomial time [14, 6]:

$$\widehat{x} = \arg\min \|x\|_1 \text{ subject to } Tx = y \tag{6}$$

These can now be solved using standard techniques like linear programming, homotopy [15] etc.

Classification using Sparse Representation

The solution to (6) provides a sparse representation of the test vector y in terms of the columns of the dictionary matrix T. In practice, (3) might be corrupted due to measurement noise or occlusion. So the model can be modified as :

$$y = Tx_0 + z \tag{7}$$

where x_0 is the sparse solution and z is due to the noise factor. So the new optimization problem can be written as

$$\widehat{x}_1 = \arg\min \|x\|_1 \text{ subject to } \|Tx - y\|_2 \leq \epsilon \tag{8}$$

where $\|z\|_2 < \epsilon$. For each class i define $\delta_i : \mathbb{R}^N \rightarrow \mathbb{R}^N$ as the characteristic function that selects the coefficients of i^{th} class only. i.e, $\delta_i(x)$ contains the coefficients of x corresponding to the i^{th} class only. Define $r_i(y) = \|y - T\delta_i(x)\|_2$ as the reconstruction residual of y w.r.t the i^{th} class. Using this function, the test vector is reconstructed w.r.t each class. Finally the identity of y is determined by the class that gives the minimal reconstruction residual.

3. SUBSPACE METHODS FOR SRC

Subspace based face recognition methods have had significant impact in the recent past. Usually the mn dimension spaces of face vectors are too difficult to handle. The most common way to handle this curse of dimension is to reduce the dimension to a level which can be comfortably handled. Principal Component Analysis [7], Linear Discriminant Analysis [8] and Locality Preserving Projections [9] are the extensively used dimension reduction techniques in face recognition.

Various dimension reduction techniques which are suitable for Sparse Representation based Classification (SRC) have already appeared in literature. Random projection [10] [11] is one of the well known method for SRC. Lu [1] proposed an optimal projection for SRC (OPSRC) and has proved superior to Principal Component Analysis(PCA), Linear Discriminant Analysis (LDA), Sparsity Preserving Projection (SPP) and Sparse Representation Classifier Steered Discriminative Projection (SRCDP). This is a supervised dimension reduction technique. The classification criterion for SRC is based on the reconstruction error corresponding to each class. The identity of a test image is the one that gives the minimum reconstruction residual. The projection matrix for OPSRC is obtained by minimizing the within class reconstruction error and simultaneously maximizing the between class reconstruction error.

OPSRC is heavy on its computational complexity. For each vector in the dictionary, it computes a within class and between class reconstruction error. For a dictionary of size $M \times N$, computational complexity for a single l_1 minimization is $\mathbb{O}(M^2N^{3/2})$. During training, the l_1 minimization problem has to be solved $\mathbb{O}(N^2)$ times. Thus the total complexity increases to $\mathbb{O}(M^2N^{5/2})$. In this paper, a method that requires minimal computation, at the same time that gives comparable discrimination as that of OPSRC is presented. This approach, motivated by the concept of Linear Discriminant Analysis, minimizes a linear objective function that minimizes the within class scatter of the data, at the same time, maximizes the between class scatter. Unlike LDA, which uses the same principle, a different objective function is used which reduces the time complexity by half. This method has a computational complexity of $\mathbb{O}(M^3)$, where $M < N$ is the dimension of test vector. Results are presented to illustrate the performance of the proposed method. A brief description of random projection, OPSRC and the proposed projection is presented.

3.1. Random Projection

In random projection, the high dimensional face data is projected on to a lower dimensional random subspace. A theorem due by Johnson and Lindenstrauss [17] states that for any set of points of size n in $\mathbb{R}^p$, there exist a linear transformation of the data into $\mathbb{R}^q$, where $q \geq \mathbb{O}(\epsilon^{-2}log(n))$ that preserves distance up to a factor of $1 \pm \epsilon$. It is computationally superior to PCA, LDA and LPP. Forming a random matrix of size $d \times M$ and projecting N vectors of dimension M to a lower dimension d takes only $\mathbb{O}(\mathbb{MN})$ computations. A condition on the matrix T that gaurentees a unique solution of (6) is called the restricted isometery property(RIP):

$$(1 - \delta)\|x\|_2 \leq \|Tx\|_2 \leq (1 + \delta)\|x\|_2 \tag{9}$$

where, δ is a small constant. In general, it is difficult to find deterministic matrices that satisfy this property. However, matrices with i.i.d Gaussian columns, Bernoulli matrices etc. have been proven to satisfy RIP with a very high probability [16]. So in this method, each face is projected on to a random subspace and this representation is used in the SRC framework.

3.2. Optimal Projection for SRC

Optimal Projection for Sparse Representation based Classification (OPSRC) [1] is a supervised dimension reduction method designed

for classification in the SRC framework. OPSRC gives a discriminative projection, such that SRC attains optimum performance in the transformed low-dimensional space.

The optimal projection P is obtained by maximizing the following objective function

$$J(P) = tr(P^T(\beta R_b - R_w)P) \tag{10}$$

where, R_b and R_w are the between class and within class reconstruction residual matrices respectively, as defined in [1]. The solution of this optimization problem are the eigen vectors corresponding to the largest d eigen vectors of the matrix $\beta R_b - R_w$. The final classification is done by doing SRC on the reduced dimension space. The computational complexity of this algorithm is $\mathbb{O}(M^2N^{5/2})$.

3.3. A new projection for SRC

The amount of computation involved in computing the optimal projection for SRC is very high. For each column of the dictionary matrix, a set of sparse coefficients needs to be computed. This drastically increases the computation involved in finding the projection matrix. Here, a new subspace projection is suggested, which is computationally efficient and achieves comparative performance to that of OPSRC. We define a linear function, similar to OPSRC, except that instead of reconstruction residuals, we use the scatter matrix defined in LDA. The objective function is:

$$\arg\max_p p^T(\alpha S_b - \beta S_w)p; \quad \alpha, \beta > 0 \tag{11}$$

$$\text{subject to } p^T p = 1$$

where, α, β are weighting parameters and S_b and S_w are the between class and within class scatter matrix as defined in (12) and (13):

$$S_b = \sum_{i=1}^{c} n_i(x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T \tag{12}$$

$$S_w = \sum_{i=1}^{c}(\sum_{j=1}^{n_i}(x_j^{(i)} - x^{(i)})(x_j^{(i)} - x^{(i)})^T) \tag{13}$$

where, $x^{(i)}$ is the mean of the i^{th} class, $x_j^{(i)}$ is the j^{th} sample of the i^{th} class. $\bar{x}$ is the global mean of the entire dataset, c is the number of distinct classes and n_i is the number of training images in the i^{th} class.

To solve the optimization problem in (11), we define Lagrange multiplier as follows:

$$\mathbb{L}(p, \lambda) = p^T(\alpha S_b - \beta S_w)p + \lambda(1 - p^T p) \tag{14}$$

Equating the partial derivative of $\mathbb{L}(p, \lambda)$ with respect to p to zero, we get:

$$\frac{\partial \mathbb{L}}{\partial p} = 2(\alpha S_b - \beta S_w)p - 2\lambda p = 0$$
$$\Rightarrow (\alpha S_b - \beta S_w)p = \lambda p$$

Thus the solution are the eigenvectors corresponding to the leading d eigenvectors $\{p_1 \dots p_d\}$ of the matrix $(\alpha S_b - \beta S_w)$. Thus $P = [p_1 \dots p_d]$ is the required projection. It can be seen that there is only one simple eigenvalue problem involved. Solution of LDA includes solution of generalized eigenvalue problem, which involves twice the computation as that of this. In PCA, the quadratic form corresponding to the total scatter matrix of the data is maximized subject to unit norm constraint:

$$\arg\max_p p^T S_t p \tag{15}$$

Table 1. Time Taken in seconds for calculating Projection Matrix

	Time in sec		
Data Size	Random Projection	OPSRC	Proposed Method
1024×90	0.0018	1791.9	0.54
1024×160	0.0092	8785.4	1.32
1024×245	0.0045	12856.7	1.74

$$\text{subject to } p^T p = 1$$

The total scatter S_t is the sum of within class S_w and between class S_b scatter matrix. Hence effectively it maximises the within class component of the scatter as well. In the proposed method, a weighted difference of between class scatter and within class scatter matrix is maximized (11). This maximises the between class scatter, at the same time minimizes the within class scatter. This enhances the discriminatory power of the projection. The results presented in the next section shows that the proposed method achieves comparable performance, to that of OPSRC, with significant reduction in computation. This method has a computational complexity of $\mathbb{O}(M^3)$. Unlike OPSRC, which uses the reconstruction residual for obtaining the projection, the proposed method uses the scatter matrix. Computing the within class reconstruction residual involves solving the l_1 minimization problem of the form (7) $\mathbb{O}(N)$ times, where N is the total number of dictionary elements. Computing the between class reconstruction residual involves solving the l_1 minimization problem $\mathbb{O}(N^2)$ times. Thus the total complexity of OPSRC algorithm increases to $\mathbb{O}(M^2N^{5/2})$.

4. EXPERIMENTAL RESULTS

The performance of the proposed method is bench marked against random projection and OPSRC. The results are presented on YALE database[18], AT&T database[19], TEXAS database[20][21]. YALE database exhibit high degree of variation in illumination. AT&T database exhibit variation in lighting, expression, slight pose variation and other variation like with/without glasses,open/closed eyes etc. TEXAS database exhibit expression variations. In each case, the face image was resized to 32×32 resolution, i.e a vector of dimension 1024. This is reduced to smaller dimensions varying from 10 to 100 using the projection methods discussed above. The classification percentage is recorded against various reduced dimension values. In all the cases the data sets were partitioned into non overlapping test and train data sets. Further, table 1 compares the time taken to build the projection matrix for random projection, OPSRC and the proposed method for various data sizes. Random projection is the fastest, since the it only needs to create a random matrix. The time taken for OPSRC is much larger compared to the other two.

YALE-B Face Database

The Yale-B face database [18] consists of 64 distinct illumination patterns of 10 subjects. Figure 3(a) shows sample images from Yale-B database. Five images of each subject with minimal illumination variation was used for training and the remaining was used for testing. Test images were randomly chosen from the remaining 59 samples for each subject. The performance is averaged over 20 runs and is presented in table 2. Figure 4 gives a graphical plot of the same. It can be seen that the proposed algorithm performs better than OPSRC and random projections.

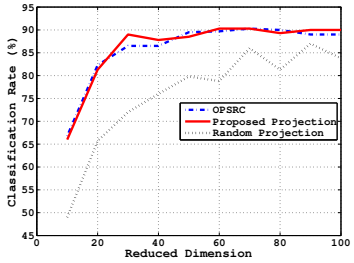

Fig. 1. Classification using Random Projection,OPSRC and proposed method for SRC AT&T Database

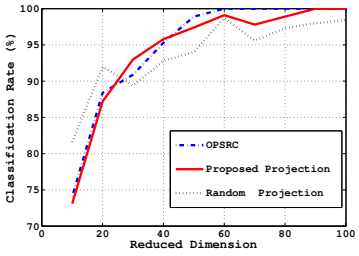

Fig. 2. Classification using Random Projection,OPSRC and proposed method for SRC TEXAS Database

(a) (b)

Fig. 3. (a) Sample images from Yale Database B (b) Sample images from AT&T database

AT&T Database

AT&T (ORL) database [19] has ten different images of 40 distinct subjects. The images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). Figure 3(b) shows some images from AT&T database. For each subject, 4 images were used for training and the remaining for testing. Figure 1 and table 3 show the performance of the proposed algorithm against random projection and OPSRC. The proposed algorithm gives a near optimal performance. At certain reduced dimensions, the proposed algorithm gives better classification performance than OPSRC.

TEXAS Database

The proposed algorithm was tried on the greyscale images of TEXAS 3D Face database [20] [21]. Though TEXAS database has images of 118 distinct subjects, only 18 subjects were used in this experiment. This is because, out of these 118 subjects, only 18 subjects have got sufficient number of distinct samples for testing and training. Five samples per subject were used for training. Table 4 shows the performance of the algorithm. Figure 2 gives a graphical representation of the same. For this database, the proposed algorithm gave comparable performance with OPSRC. In this database, there is not much illumination, pose variations and other noises. Thus the degree of scatter is less compared to other databases. Due to this reason, OPSRC tends to perform slightly better.

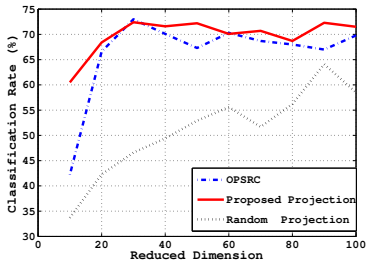

Fig. 4. Classification using Random Projection,OPSRC and proposed method for SRC Yale-B Database

Table 2. Classification results for Yale-B Database

	Reduced Dimension			
Subspace Method	40	60	80	100
OPSRC	70.1	70.3	68.0	69.8
Rand, Projection	49.4	55.6	56.2	58.8
Proposed Projection	71.6	70.1	68.7	71.5

Table 3. Classification results for AT&T Database

	Reduced Dimension			
Subspace Method	40	60	80	100
OPSRC	86.5	89.8	90.0	89.0
Rand, Projection	76	78.8	81.3	83.8
Proposed Projection	87.8	90.8	89.3	90.0

Table 4. Classification results for TEXAS database

	Reduced Dimension			
Subspace Method	40	60	80	100
OPSRC	95.3	100	100	100
Rand. Projection	92.8	98.7	97.2	98.4
Proposed Projection	95.8	99.1	98.9	100

5. CONCLUSION

Sparse representation based classification for face recognition has proven to outperform conventional face recognition techniques. However, the curse of dimension still remained a challenge for SRC. Various projection methods have been proposed to reduce the dimension of the test vector for SRC framework. OPSRC is supposed to give an optimal projection that suites SRC framework. In this paper, a new projection is introduced that gives a near optimal projection for SRC. Experimental results shows that the proposed algorithm gave comparable performance as that of OPSRC, with a much reduced computational complexity. Since the proposed method uses the data scatter matrices, it was found to perform better when there is variation in the data. The results for Yale database justifies this. The amount of computation associated with the proposed method is also much less compared to OPSRC.

6. REFERENCES

[1] Can-Yi Lu, "Optimized projection for sparse representation based classification," in *Proceedings of the 7th international conference on Advanced Intelligent Computing*, 2011, ICIC'11, pp. 83–90.

[2] W.Zhao, R.Chellappa, and A.Rosenfeld, "Face recognition: a literature survey," *ACM Computer surveys*, vol. 35, pp. 399–458, 2003.

[3] Ashok Samal and Prasana A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65 – 77, 1992.

[4] E.J.Candes, J.Romburg, and T.Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and Applied Mathematics*, vol. 59, pp. 1207–1223, 2006.

[5] E.J.Candes and T.Tao, "Near-optimal signal recovery from random projections:universal encoding strategies?," *IEEE trans on Information theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[6] J.Wright, A.Yang, A.Ganesh, S.Sastry, and Y.Ma, "Robust face recogition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[7] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 209, pp. 237–260, 1998.

[8] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711 –720, jul 1997.

[9] Xiaofei He, Shuicheng Yan, Yuxiao Hu, P. Niyogi, and Hong-Jiang Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328 –340, march 2005.

[10] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 245–250.

[11] P. Sanguansat, "Two-dimensional random projection for face recognition," in *Pervasive Computing Signal Processing and Applications (PCSPA), 2010 First International Conference on*, sept. 2010, pp. 1107 –1110.

[12] J. Ho, M.H.Yang, J. Lim, K-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proceedings of the 2003 IEEE computer society conference on Computer vision and pattern recognition*, 2003, pp. 11–18.

[13] David L. Donoho and Michael Elad, ," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–202, 2003.

[14] Donoho.D, "For Most Large Undetermined systems of Linear Equations the minimal L1 solution is also the sparsest solution," *Comm. Pure and Applied Math*, vol. 59, no. 6, pp. 797–829, 2006.

[15] Michael Elad, *Sparse and Redundant Representations:From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[16] R. Ward, "Compressed sensing with cross validation.," *IEEE Transactions on Information Theory*, , no. 12, pp. 5773 – 5782, 2009.

[17] S. Dasgupta and A. Gupta, "An elementary proof of the Johnson-Lindenstrauss lemma," 1999.

[18] A. S. Georghiades, P.N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[19] Olivetti Research Laboratories(ORL) face database, "www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html," .

[20] S. Gupta, M. K. Markey, and A. C. Bovik, "Anthropometric 3d face recognition," *International journal of Computer Vision*, vol. 90, no. 3, pp. 331–349, 2010.

[21] S. Gupta, K.R Castleman, M. K. Markey, and A. C. Bovik, "Texas 3d face recognition database," *IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 97–1, 2010.