Sparse ℓ_2-norm Regularized Regression for Face Recognition

Ahmad J. Qudaimat[1,2] and Hasan Demirel[2]

[1] *Electrical Engineering Departement, Palestine Polytechnic University (PPU), Hebron, Palestine*

[2] *Electrical and Electronic Engineering Departement, Eastern Mediterranean University (EMU), Famagusta, North Cyprus*

Keywords: Sparsifying Transform, Face Recognition, Dictionary Learning, Transform Learning.

Abstract: In this paper, a new ℓ_2-norm regularized regression based face recognition method is proposed, with ℓ_0-norm constraint to ensure sparse projection. The proposed method aims to create a transformation matrix that transform the images to sparse vectors with positions of nonzero coefficients depending on the image class. The classification of a new image is a simple process that only depends on calculating the norm of vectors to decide the class of the image. The experimental results on benchmark face databases show that the new method is comparable and sometimes superior to alternative projection based methods published in the field of face recognition.

1 INTRODUCTION

Face recognition is one of the most challenging applications of pattern recognition. It has drawn the attention of researchers for decades where many algorithms have been developed. Among the numerous number of proposed algorithms, regression-based methods with sparse representation have shown remarkable results.

In the field of pattern recognition and feature extraction, least-square regression algorithms are known by thier simplicity and effectivness. Examples of the traditional featrure extraction algorithms are principal component analysis (PCA) (Swets and Weng, 1996) which is not an optimal solution for image recognition since it based on signal reconstruction. PCA has been modified many times, for example, regression methods were used to develop sparse PCA (Zou et al., 2006) . Another example is linear Fisher discriminant analysis (LFDA) (Belhumeur et al., 1997). The idea of LFDA depends on creating a projection matrix that maximizes the ratio of intra-class scatter and inter-class scatter.

Many other regression based methods that use $\ell_{2,1}-$norm regularization were recently developed including (Ma et al., 2012), (Ma et al., 2013) and (Lai et al., 2017). In (Gu et al., 2011), Gu et al proposed the joint feature selection and subspace learning (FSSL) which was based on locality-preserving projection (LPP) (He et al., 2005). Sparsifying transform learning method has been proposed in (Qudaimat and Demirel, 2018) for face image classification (STLC).

The main contributions of this paper can be stated as follows

- A novel regression based image classification method is proposed. The method uses ℓ_2-norm regularized objective function to prevent overlapping of nonzero coefficients that belong to different classes.
- The proposed method transform images to sparse vectors using ℓ_0-norm constraint.
- Simulation results verifies competitive and superior performance of the proposed method over the alternative projection based methods.

The remaining sections of the paper are organized as follows: Section 2 discusses the details of the proposed method. Simulations and experiments are discussed in Section 3. Finally, Section 4 summarizes and concludes the proposed method.

2 PROPOSED METHOD

2.1 Problem Formulation and Objective Function

Given n training face images for K persons. In matrix form, the n_k images from k^{th} class can be represented as

$$Y_k = [y_{k1}, y_{k2}, \cdots, y_{kn_k}] \in \mathbb{R}^{N\times n_k}, \ \ k = 1,2,...,K$$

Qudaimat, A. and Demirel, H.
Sparse l2-norm Regularized Regression for Face Recognition.
DOI: 10.5220/0007355104530458
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 453-458
ISBN: 978-989-758-351-3

Let $Y = [Y_1, Y_2, \cdots, Y_K] \in \mathbb{R}^{N \times n}$ denotes the matrix representation of all subjects.

To train the dictionary W, we have formulated the following problem

$$\begin{aligned}
\text{(P1)} \min_{W,X} \quad & \|WY - X\|_F^2 \\
& - \lambda \log \Big[\sum_{i=1}^{K} \sum_{\substack{j=1 \\ i \neq j}}^{K} \|W_iY_i - W_jY_j\|_F^2 \Big] \qquad (1) \\
\text{s.t.} \quad & \|X_i\|_0 \leq s_i \ \forall i
\end{aligned}$$

where λ is a regularization parameter. W_i is a submatrix of W such that $W = [W_1^T, W_2^T, \cdots, W_K^T]^T$. For simplicity, we assume here that each submatrix W_i is a representation of consecutive number of rows in W. But in the real implementation of the algorithm, W_i represent distributed rows in W where we keep the largerst s coefficients in each column as explained in the next section.

The rational of this objective function is to increase the inner product between the transformations of the images that belong to the same class and to decrease the inner product between the transformations of the images that belong to different class. This claim becomes obvious when we expand second term in the objective function as $\|W_iY_i - W_jY_j\|_F^2 = \|W_iY_i\|_F^2 + \|W_jY_j\|_F^2 - 2(W_iY_i)^T(W_jY_j)$. In other words, ideally, W will transform images to sparse vectors where the positions of nonzero coefficients for image transformation will not overlap with image transformation of other classes.

2.2 Solution Procedure

The solution of the nonconvex optimization problem (P1) can be achieved iteratively by alternating between two steps; sparse coding and dictionary update.

step 1) sparse coding: solve (P1) by updating the vector X while keeping the dictionary W fixed. The given problem turns to the following equation

$$\min_X \|WY - X\|_F^2, \ s.t. \ \|X_i\|_0 \leq s \ \forall i \qquad (2)$$

The solution of the above equation for X can be found by zeroing all except the largest s coefficients in each column of matrix WY. Indices of these coefficients are kept in a database to be used in classification process.

step 2) dictionary update: solve (P1) by updating W while keeping X fixed. In this step (P1) is switched to following minimization problem

$$\min_W \quad \|WY - X\|_F^2 - \lambda \log \Big[\sum_{i=1}^{K} \sum_{\substack{j=1 \\ i \neq j}}^{K} \|W_iY_i - W_jY_j\|_F^2 \Big] \qquad (3)$$

optimization problem (3) has a convex differentiable objective function in W. Minimizing this objective function can be achieved by computing its derivative with respect to W and then solve for W that makes the gradient zero. The gradient of the first part is

$$\nabla_W \|WY - X\|_F^2 = 2WYY^T - 2XY^T \qquad (4)$$

To find the derivative of the second part of the objective function, consider

$$g(W) = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ i \neq j}}^{K} \|W_iY_i - W_jY_j\|_F^2 \qquad (5)$$

and

$$f(W) = \log\big[g(W)\big] \qquad (6)$$

Now, the derivative of $f(W)$ can be computed using the partial derivative $\frac{\partial g}{\partial W_k}$ as follows

$$\nabla_W f = \frac{1}{g(W)} \Big[\frac{\partial g}{\partial W_1}, \frac{\partial g}{\partial W_2}, \ldots, \frac{\partial g}{\partial W_K} \Big]^T \qquad (7)$$

To compute $\frac{\partial g}{\partial W_k}$, We first expand (5) to find the terms that includes W_k

$$\begin{aligned}
g = & \sum_{\substack{j=1 \\ i \neq j}}^{K} \|W_kY_k - W_jY_j\|_F^2 + \sum_{\substack{i=1 \\ i \neq j}}^{K} \|W_iY_i - W_kY_k\|_F^2 \\
& + \sum_{\substack{i=1 \\ i \neq k}}^{K} \sum_{\substack{j=1 \\ j \neq k \\ i \neq j}}^{K} \|W_iY_i - W_jY_j\|_F^2 \\
& = 2 \sum_{\substack{j=1 \\ j \neq k}}^{K} \|W_kY_k - W_jY_j\|_F^2 + \sum_{\substack{i=1 \\ i \neq k}}^{K} \sum_{\substack{j=1 \\ j \neq k \\ i \neq j}}^{K} \|W_iY_i - W_jY_j\|_F^2
\end{aligned} \qquad (8)$$

The second part of this equation does not depend on W_k, so its derivative is zero. Hence

$$\begin{aligned}
\frac{\partial g}{\partial w_k} & = 4 \sum_{\substack{j=1 \\ j \neq k}}^{K} (W_kY_k - W_jY_j)Y_K^T \\
& = 4 \big[(K-1)W_kY_k - \sum_{\substack{j=1 \\ j \neq k}}^{K} W_jY_j \big] Y_K^T
\end{aligned} \qquad (9)$$

To find the minimum value of the objective function we solve the following equation for every W_k

$$\begin{aligned}
& W_k \big[2YY^T - 4\lambda(K-1)Y_kY_K^T \big] - 2XY^T \\
& + 4\lambda \Big(\sum_{\substack{j=1 \\ j \neq k}}^{K} W_jY_j \Big) Y_K^T = 0
\end{aligned} \qquad (10)$$

Finally, we obtain the following closed form solution for each W_k

$$W_k = \left[2YY^T - 4\lambda(K-1)Y_k Y_K^T\right]^{-1}\left[2XY^T - 4\lambda\Big(\sum_{\substack{j=1\\ j\neq k}}^{K} W_j Y_j\Big) Y_K^T\right] \tag{11}$$

2.3 Classification

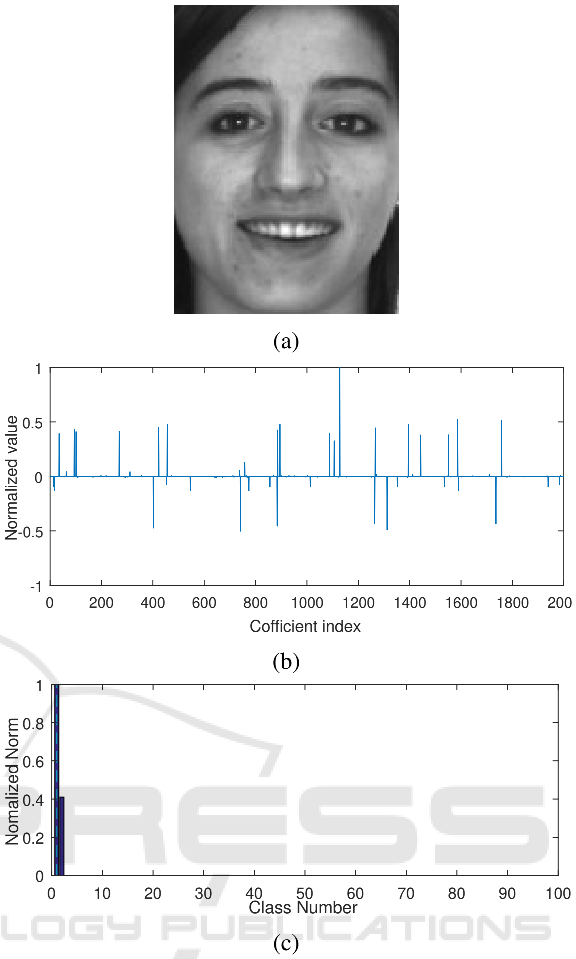
Given a new face image y_{new} to be classified to one of the classes, a new test image will be transformed using the dictionary W as

$$x = W y_{new} \tag{12}$$

Consider a function $\delta_k(x) : \mathbb{R}^L \rightarrow \mathbb{R}^s$ for each class k, which only selects the coefficients in x that correspond to class k. Then based on the maximum ℓ_2-norm of the selected coefficients, the test image y will be classified to its class as

$$class(y_{new}) = \underset{k}{\text{argmax}} \, \|\delta_k(x)\|_2^2 \tag{13}$$

Figure 1 shows an example of image transformation. In this example, the test image in figure 1a is selected from the first class of AR database. The normalized sparse vector x that is obtained is shown in figure 1b. Figure 1c shows the normalized norms of coefficeints of vector x for each class, it is obvious that the first class has the maximum norm.


(a)

(b)

(c)

Figure 1: Example of an image transformation (a) Test image image (b) The sparse coefficient vector x. (c) Norm of coefficeints for each subject.

3 EXPERIMENTAL VALIDATION

To show the performance and validity of the proposed algorithm, we test our algorithm with the benchmark face databases ORL (Samaria and Harter, 1994), AR (Martinez and Benavente, 1998), the extended-YaleB face database (Georghiades et al., 2001) and LFW databases (Huang et al., 2007). The performance of the proposed algorithm was compared with other projection based methods in litreature, i.e., PCA, LFDA (Belhumeur et al., 1997), LPP (He et al., 2005), SPP (Qiao et al., 2010), and CRP (Yang et al., 2015). nearest neighbor classifier was used.

In the experiments on LFW database, 10-fold cross validation are used. In the experiments on other databases, l face images for each subject are randomly selected to train the dictionary and the remaining images are used for testing. The experiments are repeated for the following number of training images l={2, 6} for ORL facedatabase, l={4, 8, 16} for AR face database and l={8, 16} for extended YaleB face database.

The regularization parameter λ is set to 1×10^{-9} . The number of the iterations is set to 50. The dictionary is randomely initialized. The ratio of nonzero coefficients is selected to be 10% of the total coefficients . The simulations are carried out with Intel i77500U CPU at 2.7GHz and 2.9 GHz and 12GB memory.

3.1 ORL Face Database

ORL face database (Samaria and Harter, 1994) contains 400 facial-images captured for 40 persons, 10 sample images for each one. They were taken with variances in illumination, facial expressions and facial details. Each image was cropped to size 92×112 pixels. Simulation results listed in table 1 shows that our method achieves the highest recognition rate.

Table 1: Recognition Accuracy (%) on ORL face database.

Training samples	PCA	LFDA	LPP	SPP	CRP	Proposed method
l=2	74.6	92.6	88.5	90.1	89.8	93.1
l=6	82.4	97.8	95.7	97.4	96.6	98.3

3.2 AR Database

Asubset of AR face databse has has been randomly selected. The selected subset consist of a 100 persons for 50 women and 50 men. For each one, 26 gray scale face images of dimension 165 × 120 pixels are used for training and testing the proposed algorithm. Figure 2 shows sample face images of one person.

Simulation results are shown in table 2. As shown in the table, the proposed method achieves the highest rate among all methods. The recognition rates for the proposed method and CRP are very close when the number of trainig images is 16.

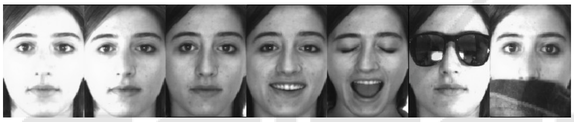


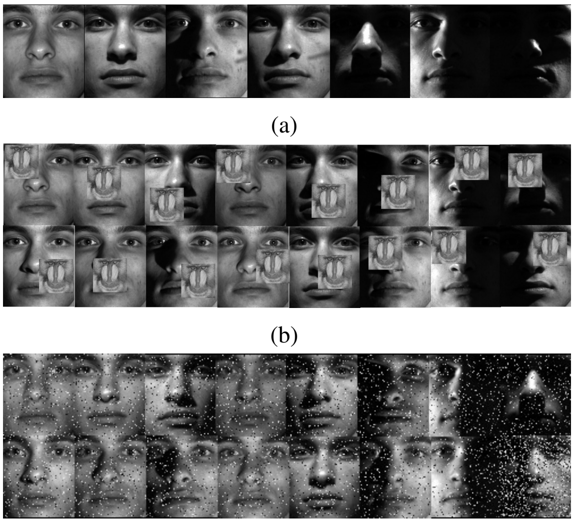
Figure 2: Sample images of AR database.

Table 2: Recognition Accuracy (%) on AR face database.

Training samples	PCA	LFDA	LPP	SPP	CRP	Proposed method
l=4	55.6	69.7	67.1	66.2	74.3	75.1
l=8	61.7	78.6	69.4	79.6	79.5	81.4
l=16	80.7	94.7	94.4	97.2	98.0	98.1

3.3 Extended Yale B Face Database

Extended YaleB (Georghiades et al., 2001) face Database comprises 38 persons with 2414 frontal person's face images. They were captured from different angles under different ligthining conditions. The images were cropped to the size 192 × 168 pixels. Each person has about 60 samples. 30 image samples are used for training and 30 image samples for testing. In these experiments, PCA was used to reduce the feature dimensions to 120. Figure 3a shows subset of face images of one person. table 3 shows the simulation results for different number of randomly selected training images.



(a)

(b)

(c)

Figure 3: Sample images of Extended Yale B databasse of (a) Samples for one peron. (b) Samples with 20% occlusion. (c) Samples with 20% corruption.

The proposed method is also tested under occlusion and corruption as shown in figure 3b and figure 3c, respectively. Table 4 shows the recognition rates under 20% occlusion, where babbon image is added to the original images, as shown in figure 3b. Results of experiments under 20% corruption are shown in table 5.

All simulation results on Extended Yale B Face Database in tables 3, 4 and 5 show the superiority of the proposed method over other methods. The proposed method achieves 97.3% recognition rate at $l = 16$. Even though the performance degrade to 90.7% in case of occluded and and to 91.6% in case of corrupted images, The proposed method still have the highest recognition rate compared to other methods under the same conditions.

Table 3: Recognition Accuracy (%) on YaleB database.

Training samples	PCA	LFDA	LPP	SPP	CRP	Proposed method
l=8	63.6	78.2	70.3	81.4	80.6	82.2
l=16	72.3	95.4	95.3	97.2	96.3	97.3

Table 4: Recognition Accuracy (%) on Extended YaleB face database with 20% block occlusion.

Training samples	PCA	LFDA	LPP	SPP	CRP	Proposed method
l=8	52.4	60.7	69.4	69.8	71.6	71.1
l=16	60.2	85.9	86.6	82.3	90.4	90.7

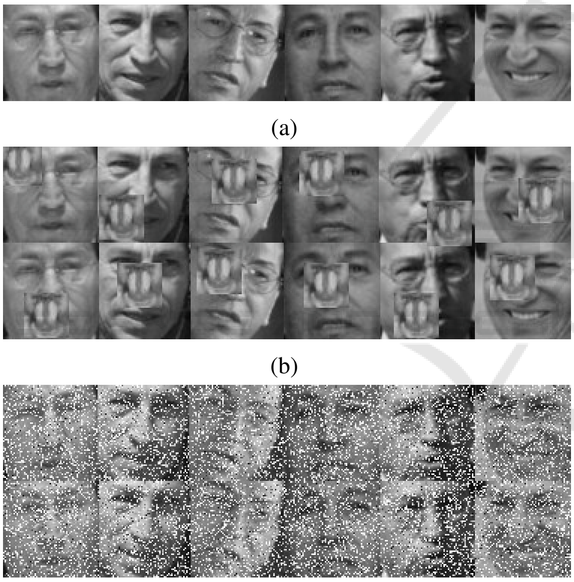
Table 5: Recognition Accuracy (%) on Extended YaleB face database with 20% corruption.

Training samples	PCA	LFDA	LPP	SPP	CRP	Proposed method
l=8	53.6	73.8	71.8	72.8	72.3	73.1
l=16	64.3	84.1	83.9	82.3	91.1	91.6

3.4 LFW Database

For this database, 100 subjects of LFWa database (Liu et al., 2015) were used. For each person, 6 images are selected. The images are cropped to eliminate the background, and resized to 64 × 64. Samples of one subject are shown in figure 4.

As in experiments with the privious databases, our method gets the highest recognition rates among all methods with LFW database as shown in table 6.



(a)

(b)

(c)

Figure 4: Sample images of LFW databasse of (a) Samples for one peron. (b) Samples with 20% occlusion . (c) Samples with 20% corruption.

Table 6: Recognition Accuracy (%) on LFW database.

Database	PCA	LFDA	LPP	SPP	CRP	Proposed method
LFW	66.9	94.1	92.5	93.8	95.1	95.3
Occluded LFW	60.9	86.8	81.3	86.1	83.2	89.1
Corrupted LFW	61.5	88.7	85.8	87.4	84.1	89.4

4 CONCLUSION

A new regression based face recognition algorithm is proposed. The method uses ℓ_0-norm to transform the image into a sparse vector. It uses ℓ_2-norm regularization to prevent the overlapping of nonzero coffecients that belongs to different subjects. The proposed method is tested with different face databases. It is compared with other well known face recognition methods. The experimental results show the superiority of accuracy of the proposed method. They also show the robustness of the method under occlusion and corruption. Another advantage of the proposed method is the low computational cost, since it only contains matrix vector multiplication and norm computation to achieve the classification task.

REFERENCES

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

Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660.

Gu, Q., Li, Z., and Han, J. (2011). Joint feature selection and subspace learning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1294–1299. AAAI Press.

He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H.-J. (2005). Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Lai, Z., Xu, Y., Yang, J., Shen, L., and Zhang, D. (2017). Rotational invariant dimensionality reduction algorithms. *IEEE Transactions on Cybernetics*, 47(11):3733–3746.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 3730–3738, Washington, DC, USA. IEEE Computer Society.

Ma, Z., Nie, F., Yang, Y., Uijlings, J. R. R., Sebe, N., and Hauptmann, A. G. (2012). Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia*, 14(6):1662–1672.

Ma, Z., Yang, Y., Sebe, N., Zheng, K., and Hauptmann, A. G. (2013). Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia*, 15(7):1628–1637.

Martinez, A. and Benavente, R. (1998). The ar face database. Technical Report 24, Computer Vision Center.

Qiao, L., Chen, S., and Tan, X. (2010). Sparsity preserving projections with applications to face recognition. *Pattern Recogn.*, 43(1):331–341.

Qudaimat, A. and Demirel, H. (2018). Sparsifying transform learning for face image classification. *Electronics Letters*, 54(17):1034 – 1036.

Samaria, F. S. and Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142.

Swets, D. L. and Weng, J. J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836.

Yang, W., Wang, Z., and Sun, C. (2015). A collaborative representation based projections method for feature extraction. *Pattern Recogn.*, 48(1):20–27.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.*, 15:1–30.