L1-norm-based (2D)^{2}PCA

Fujin Zhong[1, 2]

[1] School of Computer & Information Engineering
Yibin University name of organization
Yibin, China
E-mail: fujin-zhong@163.com

[2] School of Information Science & Technology
Southwest Jiaotong University
Chengdu, China

Abstract—**Traditional bidirectional two-dimension (2D) principal component analysis ((2D)^{2}PCA-L2) is sensitive to outliers because its objective function is the least squares criterion based on L2-norm. This paper proposes a simple but effective L1-norm-based bidirectional 2D principal component analysis ((2D)^{2}PCA-L1), which jointly takes advantage of the merits of bidirectional 2D subspace learning and L1-norm-based distance criterion. Experimental results on two popular face databases show that the proposed method is more robust to outliers than several methods based on principal component analysis in the fields of data compression and object recognition.**

Keywords-bidirectional two-dimension principal component analysis; l2-norm; outliers; L1-norm; Optimization

I. INTRODUCTION

Principal component analysis (PCA) [1] is a well-known method widely used in the areas of pattern recognition, computer vision and signal processing [2] [3]. Conventional PCA tries to learn a set of projection axes that maximize the L2-norm variance of training data. In one-dimensional PCA (termed as PCA-L2), input data, even such as image, must be scanned into a long vector so that the underlying structural information is destroyed. On the other hand, PCA-L2 must compute the eigenvectors of a large size covariance matrix because of a high-dimensional input space, which is very time-consuming. To overcome those shortcomings in image understanding, PCA-L2 is extended to two-dimension PCA (2DPCA-L2) by extracting principal components directly from two-dimension image matrices [4-6]. However, the main disadvantage of 2DPCA-L2 is that it needs much more feature dimensions for image representation than PCA-L2. Zhang and Zhou proposed the method called (2D)^{2}PCA-L2 which simultaneously adopts 2DPCA-L2 in the row and column directions of training images [7].

However, PCA-L2, 2DPCA-L2 and (2D)^{2}PCA-L2 are sensitive to outliers because their distance criterions are based on L2-norm. It is well known that the L1-norm is more robust to the presence of outliers than L2-norm [8-13]. Therefore, several PCA methods based on L1-norm have been developed to support robustness to outliers [11-13]. Among them, PCA-L1 [13] is a fast, robust and rotational invariant method which learns the local optimal projection

axes by maximizing the L1-norm-based variance in the feature space. PCA-L1's optimization algorithm is intuitive, simple and easy to implement. Li et al generalizes this algorithm to propose L1-norm-based 2DPCA (2DPCA-L1) which has lower reconstruction error than PCA-L1, PCA-L2 and 2DPCA-L2 [14]. Motivated by the relationship between (2D)^{2}PCA-L2 and 2DPCA-L2, we naturally generalizes 2DPCA-L1 to (2D)^{2}PCA-L1. The proposed method extracts the local optimal projection vectors based on maximization of the L1-norm-based variance simultaneously in the row and column directions of training images. The experimental results show the robustness of our proposed method to outliers in image reconstruction and object recognition.

II. (2D)^{2}PCA-L1

Let $\mathbf{X}_i$ ($i=1, \ldots, N$) denote N training image samples where the size of $\mathbf{X}_i$ is $m \times n$. 2DPCA-L1 or alternative 2DPCA-L1 only learns in row or column direction of training image samples respectively. 2DPCA-L1 learns an optimal projection matrix $\mathbf{Q}' = [\mathbf{q}_1', \ldots, \mathbf{q}_d'] \in \mathbf{R}^{n \times d}$ from the information between rows of training images, and then projects image $\mathbf{X} \in \mathbf{R}^{m \times n}$ onto $\mathbf{Q}'$ to gain a matrix $\mathbf{Y} \in \mathbf{R}^{m \times d}$. Alternative 2DPCA-L1 learns an optimal projection matrix $\mathbf{P}' = [\mathbf{p}_1', \ldots, \mathbf{p}_k'] \in \mathbf{R}^{m \times k}$ just from the information between columns of training images, and then projects image $\mathbf{X}$ onto $\mathbf{P}'$ to gain a matrix $\mathbf{A} \in \mathbf{R}^{k \times n}$. We propose our method to project image $\mathbf{X}$ onto $\mathbf{P}'$ and $\mathbf{Q}'$ simultaneously, which is called L1-norm-based (2D)^{2}PCA or (2D)^{2}PCA-L1.

A. 2DPCA-L1

The 2DPCA-L1 method is a 2-D extension of PCA-L1. Let $\mathbf{q}' \in \mathbf{R}^{n \times 1}$ is the first projection vector which is the convergence result of iterations, so we can use $\mathbf{q}'(t)$ to denote the result of the tth iteration. Then, the projection of the original image $\mathbf{X}_i \in \mathbf{R}^{m \times n}$ onto $\mathbf{q}'$ produces the corresponding feature $\mathbf{y}_i \in \mathbf{R}^{m \times 1}$ as

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{q}' = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{im} \end{bmatrix} \mathbf{q}' \qquad (1)$$

where $\mathbf{x}_{ir} \in \mathbf{R}^{1 \times n}$ is the rth row vector of $\mathbf{X}_i$.

2DPCA-L1 aims at maximizing the L1-norm variance in low-dimensional feature space

$$J(\mathbf{q}') = \sum_{i=1}^{N} \|\mathbf{y}_i\|_1 = \sum_{i=1}^{N} \sum_{r=1}^{m} |\mathbf{x}_{ir} \mathbf{q}'| \qquad (2)$$

Here, $\|\bullet\|_1$ denotes the L1-norm of a vector or a matrix. The optimal $\mathbf{q}*$ is derived as the following:

$$\mathbf{q}* = \underset{\mathbf{q}'}{arg\ max}\, J(\mathbf{q}') \text{ subject to } \|\mathbf{q}'\|_2 = 1 \qquad (3)$$

Here, $\|\bullet\|_2$ denotes the L2-norm of a vector or a matrix. The algorithm to solve (3) can be described as follows:

1) Set $t=0$ and $\mathbf{q}'(t) = (\underset{\mathbf{x}_{ij}}{arg\ max} \|\mathbf{x}_{ir}\|_2)^T$, and rescale $\mathbf{q}'(t)$ to unit length.

2) Define one polarity function $f_{ir}(t)$ as

$$f_{ir}(t) = \begin{cases} 1, & \text{if } |\mathbf{x}_{ir}\mathbf{q}'(t)| \geq 0 \\ -1 & \text{if } |\mathbf{x}_{ir}\mathbf{q}'(t)| < 0 \end{cases} \qquad (4)$$

3) Let $\mathbf{q}'(t+1)$ be updated by

$$\mathbf{q}'(t+1) = \frac{\sum_{i=1}^{N}\sum_{r=1}^{m} f_{ir}(t)\mathbf{x}_{ir}^T}{\left\| \sum_{i=1}^{N}\sum_{r=1}^{m} f_{ir}(t)\mathbf{x}_{ir}^T \right\|_2} \qquad (5)$$

The convergence of iteration equation (5) can be justified as in [14].

4) If $J(\mathbf{q}'(t+1)) = J(\mathbf{q}'(t))$, then stop the iteration and set $\mathbf{q}* = \mathbf{q}'(t)$. Otherwise, go to Step 2.

5) Output $\mathbf{q}*$.

When the first optimal projection vector $\mathbf{q}'_1$ is attained, we can compute $\mathbf{q}'_s$, with $s>1$, by updating the training samples by

$$\mathbf{x}_{ir}^{s} = \mathbf{x}_{ir}^{s-1} - \mathbf{x}_{ir}^{s-1}\mathbf{q}'_{s-1}(\mathbf{q}'_{s-1})^T, i=1,\cdots,N, r=1,\cdots,m \quad (6)$$

It can be proven that $\mathbf{q}'_s$ is orthogonal to $\mathbf{q}'_{s-1}$ as described in [14].

B. Alternative 2DPCA-L1

We can also use 2DPCA-L1 learning in the column direction of images called alternative 2DPCA-L1. Similarly,

alternative 2DPCA-L1 aims at maximizing the L1-norm variance in low-dimensional feature space projected from the column vectors of image samples. Let $\mathbf{x}_{ic} \in \mathbf{R}^{m \times 1}$ is the cth column vector of $\mathbf{X}_i$, and $\mathbf{p}* \in \mathbf{R}^{m \times 1}$ is the optimal projection vector in the column direction, which can be computed by

$$\mathbf{p}* = \underset{\mathbf{p}'}{arg\ max}\, J(\mathbf{p}') = \underset{\mathbf{p}'}{arg\ max} \sum_{i=1}^{N} \|(\mathbf{p}')^T \mathbf{X}_i\| \qquad (7)$$

$$= \underset{\mathbf{p}'}{arg\ max} \sum_{i=1}^{N}\sum_{c=1}^{n} |(\mathbf{p}')^T \mathbf{x}_{ic}| \text{ subject to } \|\mathbf{p}'\|_2 = 1$$

The algorithm to solve (7) is similar to the solution of (3). Similarly, when the first optimal projection vector $\mathbf{p}'_1$ is attained, we can compute $\mathbf{p}'_s$, with $s>1$, by updating the training samples by

$$\mathbf{x}_{ic}^{s} = \mathbf{x}_{ic}^{s-1} - (\mathbf{p}'_{s-1})^T \mathbf{x}_{ic}^{s-1}\mathbf{p}'_{s-1}, i=1,\cdots,N, c=1,\cdots,n \quad (8)$$

Similarly, $\mathbf{p}'_s$ is orthogonal to $\mathbf{p}'_{s-1}$.

C. (2D)²PCA-L1

(2D)²PCA-L1 aims to project image $\mathbf{X}$ onto $\mathbf{P}'$ and $\mathbf{Q}'$ in the row and column directions of training images simultaneously. Firstly, we can respectively gain the projection matrices $\mathbf{P}'$ and $\mathbf{Q}'$ as described in aforementioned two subsections. Then, any original image $\mathbf{X}$ can be projected to yield a k by d matrix $\mathbf{Z}'$ as

$$\mathbf{Z}' = (\mathbf{P}')^T \mathbf{X}\mathbf{Q}' \qquad (9)$$

The matrix $\mathbf{Z}'$ is also called the feature matrix of the original image $\mathbf{X}$ in image representation. Similarly, the image $\mathbf{X}$ can be reconstructed by

$$\hat{\mathbf{X}} = \mathbf{P}'\mathbf{Z}'(\mathbf{Q}')^T \qquad (10)$$

The reconstruction error can be computed by

$$e = \|\mathbf{X} - \hat{\mathbf{X}}\|_2 = \|\mathbf{X} - \mathbf{P}'\mathbf{Z}'(\mathbf{Q}')^T\|_2 \qquad (11)$$

In object recognition, each training image samples $\mathbf{X}_i$ ($i=1, \ldots, N$) is firstly projected onto both $\mathbf{P}'$ and $\mathbf{Q}'$ simultaneously to obtain the feature matrix $\mathbf{Z}'_i$ ($i=1, \ldots, N$). Secondly, let $\mathbf{X}$ be a given image sample for recognition and then get its feature matrix $\mathbf{Z}'$ similarly by projecting it onto both $\mathbf{P}'$ and $\mathbf{Q}'$ simultaneously. Lastly, an Euclidean distance based nearest neighbor classifier [15] is used for object classification. Here, the distance between $\mathbf{Z}'$ and $\mathbf{Z}'_i$ is defined as following

$$d(\mathbf{Z}', \mathbf{Z}'_i) = \|\mathbf{Z}' - \mathbf{Z}'_i\|_2 = \sqrt{\sum_{s=1}^{k}\sum_{t=1}^{d}(\mathbf{z}'^{(s,t)} - \mathbf{z}_i'^{(s,t)})^2} \quad (17)$$

III. EXPERIMENTAL RESULTS

In this section, we experimentally compare our proposed $(2D)^2$PCA-L1 with $(2D)^2$PCA-L2, 2DPCA-L2, 2DPCA-L1, PCA-L2 and PCA-L1 on two well-known face databases: ORL face database [16] and Yale face database [17]. We note that $(2D)^2$PCA-L1, $(2D)^2$PCA-L2, 2DPCA-L1 and 2DPCA-L2 are image-as-matrix methods, while PCA-L1 and PCA-L2 are image-as-vector methods. Therefore, the performances of those methods are compared on the basis of the same total size of feature vectors or matrices for the fairness. Generally, subspace learning algorithms can be used for data compression, object recognition, image retrieval, and other applications [14]. In this paper, Yale database is used to test the performance of image compression and ORL database is used to test the performance of object recognition when random outliers are present in the training dataset. In addition, all experiments are executed on a computer system of Intel T2350 1.86GHz and 1GB RAM with Matlab 7.9.

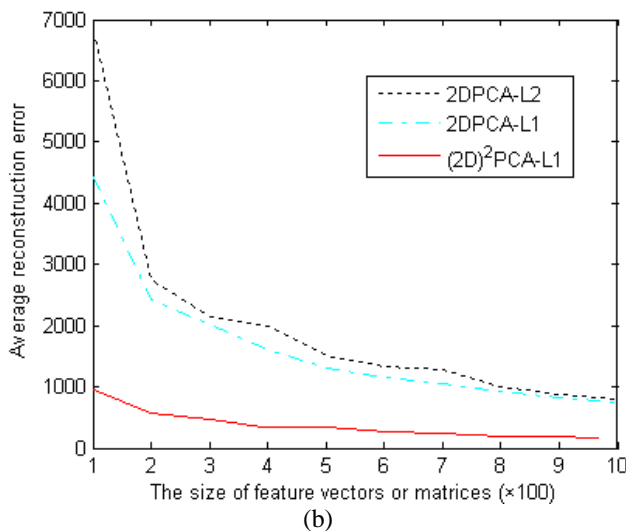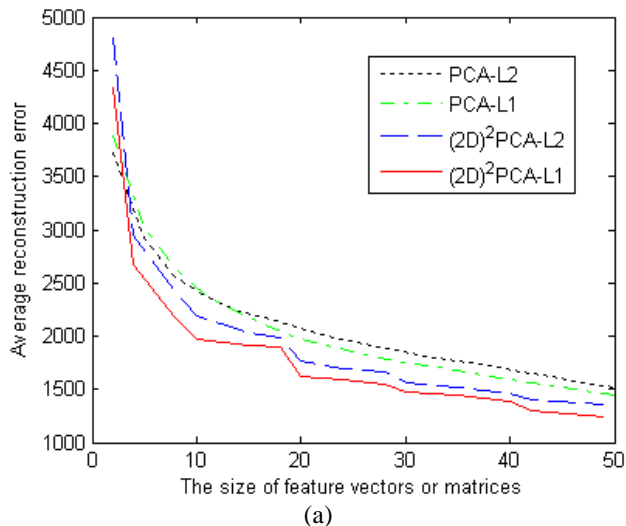### A. Face reconstruction on Yale face database

Figure 1. Average reconstruction error versus the size of feature vectors or matrices on Yale database.

In order to compare the performances of different methods in terms of image compression, we give the average reconstructed error of inliers for objective evaluation when outliers are present in the training data as in [13] and [14]. The experiments are carried on Yale face database which includes 165 grayscale images of 15 individuals. All images are normalized so that the ground truth positions of the two eyes and mouth are the same. Then, every image is cropped and scaled to 100×100 size. In addition, among 165 images, 20 percent were randomly selected as outliers by being occluded with a random rectangular noise. The rectangle consists of back or white dots in random distribution, its location in image is random and its sizes range from 15×10 to 60×60. For fairness, the results of all methods must be on the basis of the same feature size. Among these six methods, 2DPCA-L2 and 2DPCA-L1 are special because the size of their features must be the multiple of 100. Therefore, we partition the comparing results into two groups. Fig. 1 (a) shows the average reconstruction errors of the following methods: PCA-L2, PCA-L1, $(2D)^2$PCA-L2 and $(2D)^2$PCA-L1 versus the feature size. And, Fig. 1 (b) shows the average reconstruction errors of the following methods: 2DPCA-L2, 2DPCA-L1 and $(2D)^2$PCA-L1 versus the feature size. Obviously, it can be found that the average reconstruction error of $(2D)^2$PCA-L1 is lower than those of the other five methods. In image compression, the results show that $(2D)^2$PCA-L1 is the most robust to the outliers.

B. Face recognition on ORL face database

Figure 2. Some samples with occlusion in ORL database.

TABLE I.   AVERAGE OPTIMAL RECOGNITION RATES (%) AND ITS STANDARD DEVIATION ON ORL DATABASE

Methods	2 Training Samples	3 Training Samples	4 Training Samples	5 Training Samples
PCA-L2	75.75±2.94	82.79±2.82	86.71±2.11	90.55±1.48
PCA-L1	74.38±3.05	83.86±1.50	86.58±2.85	91.05±1.54
2DPCA-L2	82.46±3.26	88.17±2.20	91.86±1.72	93.55±1.62
2DPCA-L1	82.75±3.18	88.60±2.18	92.06±1.59	93.65±1.49
(2D)2PCA-L2	83.60±3.34	89.24±2.12	92.92±1.67	95.10±1.43
(2D)2PCA-L1	83.79±3.19	89.55±2.09	93.11±1.84	95.20±1.53

This part compares the performance of these six methods on ORL face database in the field of object recognition. The ORL database includes 400 images of 40 persons, every person providing 10 different images with the size of 112×92. All images are gray with 256 levels and some images were taken at different time [18]. In our experiment, the grayscale images are not preprocessed, except that 20 percent of these 400 images are randomly selected to be

added rectangle noise which is generated as in subsection III.A. Fig. 2 shows some samples with occlusion as outliers. We compare the optimal recognition rates of these six methods with different number of training samples. The k (=2 3, 4, 5) training samples are randomly selected from each subject's 10 images and the remaining images of every subject are used as the test data. The experiments are repeated fifteen times with each special training sample number. We record the optimal recognition rates of each method. Then, the average optimal recognition rate and its standard deviation are calculated and Table 1 gives the results. From Table 1, we can conclude that the average optimal recognition rate of $(2D)^2$PCA-L1 is slightly higher than that of $(2D)^2$PCA-L2 but significantly higher than those of the other four methods .

IV. CONLUSION

In this paper, an L1-norm-based bidirectional two-dimension principal component (termed as $(2D)^2$PCA-L1) is proposed. $(2D)^2$PCA-L1 learns the local optimal projection vectors based on maximization of the L1-norm-based variance simultaneously in the row and column directions of training images. $(2D)^2$PCA-L1 jointly takes advantage of the merits of bidirectional 2D subspace learning and L1-norm distance criterion. In the fields of data compression and object recognition, the experimental results on two popular face databases have demonstrated the efficacy of the proposed $(2D)^2$PCA-L1which is not only more robust to outliers but also uses less feature size than several algorithms based on principal component analysis.

REFERENCES

[1] I.T. Jolliffe, Principal component analysis, Springer, NewYork, 1986.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," J. Cognit. Neurosci., vol. 3 no. 1, 1991, pp. 71–86.

[3] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: a review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, 2000, pp. 4–37.

[4] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 1, 2004, pp. 131–137.

[5] H. Kong, L. Wang, E. K. Teoh, J. Wang, and R. Venkateswarlu, "Generalized 2D principal component analysis for face image representation and recognition," Neural Netw., vol. 18, no. 5/6, 2005, pp. 585–594.

[6] Y. Pang, D. Tao, Y. Yuan, and X. Li, "Binary two-dimensional PCA," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 38, no. 3, 2008, pp. 1176-1180.

[7] D. Zhang and Z. Zhou, "(2D)²PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," Neurocomputing, vol. 69, no.1-3, 2005, pp. 224–231.

[8] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," Int. J. Comput. Vis., vol. 19, no. 1, 1996, pp. 57–91.

[9] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," IEEE Trans. Biomedical Engineering, vol. 59, no. 3, 2012, pp. 653-662.

[10] Y. Pang, X. Li, and Y. Yuan, "Robust tensor analysis with L1-norm," IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 2, 2010, pp. 172–178.

[11] A. Baccini, P. Besse, and A.D. Falguerolles, A L1-Norm PCA and a heuristic approach, Ordinal and Symbolic Data Analysis, E. Diday, Y. Lechevalier, and P. Opitz, eds., pp. 359-368, Springer, 1996.

[12] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization," Proc. Int. Conf. Mach. Learn. , 2006, pp. 281–288.

[13] N. Kwak, "Principal component analysis based on L1-norm maximization," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 9, 2008, pp. 672–1680.

[14] X. L. Li, Y. W. Pang, and Y. Yuan, "L1-norm-based 2DPCA," IEEE Trans. Syts., Man, Cybern. B, Cybern., vol. 40, no. 4, 2009, pp. 1170-1175.

[15] C.J. Veeman and M.J.T. Reinders, "The nearest subclass classifier: a compromise between the nearest mean and nearest neighbor classifier," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 9, 2005, pp. 1417-1429.

[16] The ORL Face Database, Cambridge, U.K, AT&T (Olivetti) Research Laboratories. [Online]. Available: http://www.uk.research.att.com/facedatabase.html.

[17] Yale Face DB: Yale Univ.2007. [Online]. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[18] Y.F. Qi and J.S. Zhang, "(2D)2PCALDA: An efficient approach for face recognition," Applied Mathematics and Computation, vol. 213, 2009, pp. 1–7.