Mahalanobis Distance Based Non-negative Sparse Representation for Face Recognition

Yangfeng Ji, Tong Lin, Hongbin Zha
Key Laboratory of Machine Perception (Ministry of Education),
School of EECS, Peking University,
Beijing 100871, China

Abstract

Sparse representation for machine learning has been exploited in past years. Several sparse representation based classification algorithms have been developed for some applications, for example, face recognition. In this paper, we propose an improved sparse representation based classification algorithm. Firstly, for a discriminative representation, a non-negative constraint of sparse coefficient is added to sparse representation problem. Secondly, Mahalanobis distance is employed instead of Euclidean distance to measure the similarity between original data and reconstructed data. The proposed classification algorithm for face recognition has been evaluated under varying illumination and pose using standard face databases. The experimental results demonstrate that the performance of our algorithm is better than that of the up-to-date face recognition algorithm based on sparse representation.

1 Introduction

Face recognition has attracted a great deal of attention in computer vision and machine learning. After many years of research, high performance can now be achieved under controlled conditions. However when variations exist due to extrinsic factors like pose and illumination, the performance decreases remarkably. In this paper, we focus on the problem of face recognition under the conditions of varying illumination and pose. We propose an improved sparse representation based classification algorithm for face recognition. The problem of sparse representation is to search the most compact representation of data in terms of linear combination of atoms in an overcomplete dictionary. In machine learning, sparse representation is popular in feature extraction, classification, etc.

For classification, there are some works related to sparse representation. Huang et al. [5] propose a sparse representation based classification algorithm combined with linear discriminative analysis (LDA) for signal classification. The experimental results show that the proposed method outperforms the standard sparse representation and the standard discriminative methods in the case of corrupted signals. However, this algorithm can not be generalized to multi-class classification easily. Wright et al. [11] exploit a method of sparse representation for face recognition. They investigate the problem of face recognition with the theory of compressive sensing. By using the theory of compressive sensing [2, 3], they propose a new classification algorithm, named the sparse representation-based classification (SRC) algorithm and show the performance on face recognition under varying illumination and partial occlusion. However, the SRC algorithm does not involve some special factors in face recognition, such as similarity measure using more suitable distance than Euclidean distance, etc.

To improve the performance of the sparse representation based classification, we propose a non-negative sparse representation based classification algorithm using Mahalanobis distance, and its applications on face recognition. First, we address the problem of sparse representation using the constraint of non-negative sparse coefficient to obtain a discriminative representation. Second, we replace Euclidean distance with Mahalanobis distance to measure similarity between original data and reconstructed data. Then, we reformulate the problem to be an equivalent ℓ_1-regularized least square problem for obtaining its solution. In experiments on face recognition, we compare the improved algorithm with the SRC algorithm on two different face databases, all experimental results show the performance of the improved algorithm is better than that of the SRC algorithm.

2 Non-negative Sparse Representation

For a discriminative representation, we require the sparse coefficient to be non-negative. Therefore, for a given test

sample, components of coefficient indicate the contributions of training samples. Furthermore, Mahalanobis distance is employed to measure the similarity between original data and reconstructed data, instead of Euclidean distance.

2.1 Sparse representation in subspace

Generally, given n high-dimensional data points $A = \{a_1, \ldots, a_n\}$, some research on manifold learning, for instance LLE [10], has proved that these data lie on a lower dimensional manifold. Any data point $a_i \in A$ can be approximately represented by the linear combination of its neighboring data points. This kind of linear representation can be generalized to labeled data. Given data points $\{a_1, \ldots, a_n\}$ in one class, a new data point a^* in the same class can be represented as linear combination of $\{a_1, \ldots, a_n\}$,

$$a^* = \beta_1 a_1 + \ldots + \beta_n a_n. \tag{1}$$

In other words, given n training samples $\{a_1, \ldots, a_n\}$, their linear combinations span a linear subspace $\mathcal{X}$:

$$\mathcal{X} = \text{span}\{a_1, \ldots, a_n\}.$$

A new sample a^* in the same class approximately lie on this subspace.

Linear representation for labeled data can be used in pattern recognition. Given sufficient K classes training samples, a basic problem in pattern recognition is to correctly determine the class which a new coming (test) sample belongs to. We arrange the n_k training samples from the kth class as columns of a matrix $A_k = [a_{k,1}, \ldots, a_{k,n_k}] \in \mathcal{R}^{m \times n_k}$. Then, we obtain the training sample matrix $A = [A_1, \ldots, A_K]$. Under the assumption of linear representation, a test sample $y \in \mathcal{R}^m$ will approximately lie on the linear subspace spanned by training samples,

$$\begin{aligned} y \quad = \quad & \beta_{1,1} a_{1,1} + \ldots + \beta_{1,n_1} a_{1,n_1} \\ & + \ldots \\ & + \beta_{K,1} a_{K,1} + \ldots + \beta_{K,n_K} a_{K,n_K}. \end{aligned} \tag{2}$$

or, in matrix form,

$$y = Ax \quad \in \mathcal{R}^m, \tag{3}$$

where x is a coefficient vector. For accurate reconstruction of sample y in class k,

$$x = [0, \ldots, 0, \beta_{k,1}, \ldots, \beta_{k,n_k}, 0, \ldots, 0]^T \in \mathcal{R}^n.$$

If K is large, x will be sufficient sparse. However, for many practical problems, accurate reconstruction is nearly impossible.

If $m < n$, Eq.(2) is under-determined, and its solution is not unique. This motivates us to solve the following optimization problem for a sparse solution:

$$\hat{x} = \arg\min_x \|x\|_0 \quad \text{subject to} \quad Ax = y, \tag{4}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm, which counts the number of nonzero entries in a vector. However, the problem of finding the sparse solution of Eq.(4) is NP-hard, and difficult to solve.

The theory of compressive sensing [2, 3] reveals that if the solution x is sparse enough, we can solve the following convex relaxed optimization problem to obtain approximate solution:

$$\hat{x} = \arg\min_x \|x\|_1 \quad \text{subject to} \quad Ax = y \tag{5}$$

Furthermore, supposing that the observations are inaccurate, we should relax the constraint in Eq.(5) and have the following optimization problem:

$$\hat{x} = \arg\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 \leq \varepsilon \tag{6}$$

where ε is the tolerance of reconstruction error.

Regularization is one of the most popular methods to deal with constrained optimization problems, for instance, in the theory of Support Vector Machine. For Eq.(6), that is

$$\hat{x} = \arg\min_x \|x\|_1 + \gamma \|Ax - y\|_2^2. \tag{7}$$

where γ is a weight to make a trade-off between reconstruction error and sparsity in the representation.

2.2 Non-negative constraint for sparse coefficient

Sparse representation for classification is different from that for signal reconstruction. In signal processing, an original signal y should be reconstructed as accurately as possible. However, in classification, a discriminative representation is more important than reconstruction accuracy.

For a discriminative representation, we require that coefficient x should indicate contributions of all training samples to a given test sample. Therefore, we add constraint $x \geq 0$ to Eq.(7)

$$\hat{x} = \arg\min_{x: x \geq 0} \|x\|_1 + \gamma \|Ax - y\|_2^2, \tag{8}$$

where $\hat{x}$ is sparse, in which all elements are non-negative. The sparse representation from Eq.(8) avoids "negative" contribution of some training samples. In this way, for a given test sample, the similar training samples can be found from sparse representation.

2.3 Similarity measure using Mahalanobis distance

For measure the similarity between original data and reconstructed data, We employ Mahalanobis distance instead of Euclidean distance. By introducing Mahalanobis distance, we obtain a generalized distance measure for face recognition, which can embody different weights on different components of feature vector. Mahalanobis Distance has been proved as a better similarity measure than Euclidean distance, when it comes to pattern recognition problems, for instance, face recognition [9].

Given two data points $v_1, v_2 \in \mathcal{R}^m$, their Mahalanobis distance is given by:

$$\begin{aligned} d_M(v_1, v_2) &= \|v_1 - v_2\|_M \\ &= \sqrt{(v_1 - v_2)^T M (v_1 - v_2)}, \end{aligned} \quad (9)$$

where $M \in \mathcal{R}^{m \times m}$ is a positive definite matrix.

Using the definition of Mahalanobis distance, the distance between original data y and reconstructed data Ax is

$$d_M(Ax, y) = \|Ax - y\|_M = \sqrt{(Ax - y)^T M (Ax - y)}.$$

The objective function with Mahalanobis distance can be formulated as follows:

$$\hat{x} = \arg\min_{x: x \geq 0} \|x\|_1 + \gamma \|Ax - y\|_M^2. \quad (10)$$

There are three different types of positive definite M in the definition of Mahalanobis distance:

- M is any positive definite matrix.
- M is a diagonal matrix. If the type of M is diagonal, M gives different components with different weights.
- M is a scalar multiple of the identity matrix I, $M = \sigma I$, where $\sigma > 0$. If $M = \sigma I$, Eq.(10) is equivalent to Eq.(8) with regularization coefficient $\sigma\gamma$.

In this paper, the matrix M is determined by the importance of components of feature vectors. Moreover, we can ingore the correlation among different dimensions.

3 Algorithm

Using the Cholesky factorization, the problem of Mahalanobis distance based non-negative sparse representation can be solved by a standard optimization algorithm. Then, the classification algorithm is designed based on the idea of finding the minimal reconstruction error [11].

3.1 Algorithm for solving non-negative ℓ_1-regularized least square

Since M is a positive definite matrix, the Cholesky factorization of M is

$$M = L^T L, \quad (11)$$

where U is a lower triangular matrix with positive diagonal entries. From Eq.(11), the objective function in Eq.(10) can be formulated as:

$$\begin{aligned} \hat{x} &= \arg\min_{x: x \geq 0} \|x\|_1 + \gamma((Ax - y)^T L^T L (Ax - y)) \\ &= \arg\min_{x: x \geq 0} \|x\|_1 + \gamma((LAx - Ly)^T (LAx - Ly)) \\ &= \arg\min_{x: x \geq 0} \|x\|_1 + \gamma \|LAx - Ly)\|_2. \end{aligned} \quad (12)$$

Set $A' = LA$ and $y' = Ly$. Given parameter $\gamma > 0$, the problem is equal to the following problem:

$$\begin{aligned} \hat{x} &= \arg\min_{x: x \geq 0} \|x\|_1 + \gamma \|A'x - y')\|_2^2 \\ &= \arg\min_{x: x \geq 0} \lambda \|x\|_1 + \|A'x - y'\|_2^2, \end{aligned} \quad (13)$$

where $\lambda = \gamma^{-1}$. Eq.(13) is a non-negative ℓ_1-regularized least square problem, which can be solved by second-order cone programming [1, 8].

3.2 Recognition algorithm

The recognition algorithm is inspired by the SRC algorithm proposed in [11]. Given a test sample y, we first compute its sparse coefficient $\hat{x}$. Then, we determine the class of this test sample from its reconstruction error between this test sample and the training samples of class k,

$$E_k(\hat{x}) = \|A\delta_k(\hat{x}) - y\|_M, \quad (14)$$

where residual error is computed using Mahalanobis distance. For each class k, $\delta_k(x) : \mathcal{R}^n \to \mathcal{R}^n$ is the characteristic function which selects the coefficients associated with the kth class. The class $C(y)$ which test sample y belongs to is determined by

$$C(y) = \arg\min_k E_k(\hat{x}). \quad (15)$$

The whole algorithm of our method is summarized in algorithm 1.

4 Experiments

In experiments, we test our algorithm on face recognition with two face databases: the extended Yale face database B

Algorithm 1 Our algorithm

Input: Test sample y, training matrix A, parameter γ

1: Normalize the columns of A using ℓ_2 norm

2: Solve

$$\hat{x} = \arg\min_{x:x\geq 0} \|x\|_1 + \gamma\|Ax - y\|_M^2$$

using an equivalent non-negative ℓ_1-regularized least square problem

3: Compute reconstruction error $E_k(k = 1, ..., K)$:

$$E_k(\hat{x}) = \|A\delta_k(\hat{x}) - y\|_M$$

Output: $C(y)$, where $C(y) = \arg\min_k E_k(\hat{x})$

and the Sheffield face database. For feature extraction, we downsample face images to some given sizes and reshape them to be column vectors. Then we arrange all column vector of training samples to be a matrix A in Eq.(10). For a test sample, y, our algorithm and the SRC algorithm [11] are employed to obtain the results of recognition.

4.1 Illumination variant database — the extended Yale face database B

First, we evaluate the performance of our algorithm on a cropped version of the Extended Yale Face Database B [7]. There are 2,427 face images of 38 individuals in this database. All images are cropped into size of 192×168 pixels. There are different lighting conditions on each image for each subject(see Figure 1 for some examples). We randomly choose half of images in each class for training and the other images for testing. For feature extraction, we downsample all images to six different sizes: 30, 56, 72, 90, 110 and 120.

We compare the performance of our algorithm with that of the SRC algorithm [11] on different dimensions of features. As illustrated in Figure 2, our algorithm has better performance than the SRC algorithm. The best performance of our algorithm in this experiment is 96.74%, while the best performance of the SRC algorithm is 95.05%. As shown in this experiment, the classification accuracy is improved by non-negative coefficient constraint and Mahalanobis distance.

4.2 Pose variant database — the Sheffield face database

In this section, the algorithm performance is tested on the subset of Sheffield (previously UMIST) face database [4]. This subset of Sheffield face database consists of 495 face images of 18 individuals. Faces in the database cover range

Figure 1. Sample faces from different persons under different illumination in the extended Yale face database B

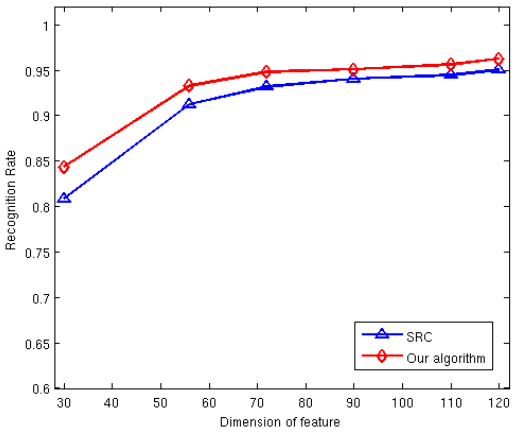

Figure 2. Comparsion of recognition rates of our algorithm and the SRC algorithm on the extended Yale face database B

of poses from frontal view to profile. All persons in this database cover a mixed range of race, sex and age. Some images from Figure 3 are examples of this database.

Features are extracted from all original images before preforming face recognition using downsampling. In this experiment, we downsample all images from size of 112×96 to different sizes, such as 6×5, 8×7, 9×8, etc. Then, all downsampling images are reshaped to column vectors as feature vectors. We divide all images to training samples and test samples: 94 training images and 401 test images. For each subject, there are about $5-9$ images from different poses for training, and the other images for testing.

We also compare the performance of our algorithm with that of the SRC algorithm on different dimensions of feature. In figure 4, we can see that our algorithm is slightly

better than the SRC algorithm. Ranging from different dimension, the best performance of our algorithm is 97.51%, compared with 96.26% of the SRC algorithm.

Note that, when the dimension of feature continues to increase, the performance of the SRC algorithm decreases remarkably. For example, when the dimension of feature vector is 81, the recognition rate of the SRC algorithm is 82.54%. The recognition rate of our algorithm is 96.51% at the same dimension.

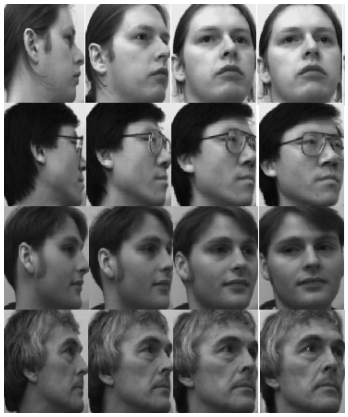

Figure 3. Sample faces from different persons under different pose in the Sheffield face database

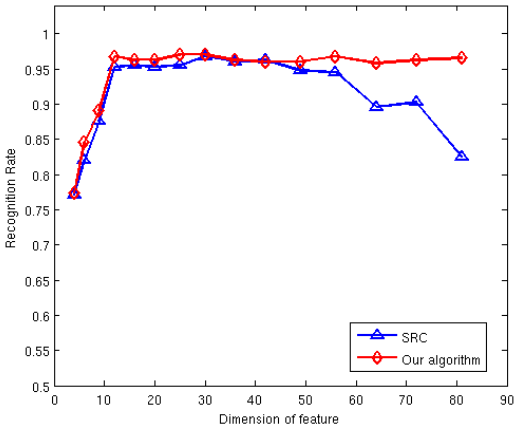

Figure 4. Comparsion of recognition rates of our algorithm and the SRC algorithm on the Sheffield face database

5 Conclusion

In this paper, we propose an improved classification algorithm based on non-negative sparse representation with Mahalanobis distance. For a discriminative representation, a non-negative constraint of sparse coefficient is added. Moreover, Mahalanobis distance is used as a measure of image similarity instead of Euclidean distance in feature space. Then, we build a connection between this problem and non-negative ℓ_1-regularized least square problem. The experimental results on face recognition show that the performance of our algorithm is better than the SRC algorithm. In the future, we will apply the framework of Mahalanobis distance based non-negative sparse representation to other fields, such as, object detection, etc.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and the helpful discussion with Wei Ma and Tao Luo. This work was partially supported by the National Science Foundation of China(NSFC) Grants 60775006, National Key Basic Research Program(NKBRP) Grant 2004CB318005, and the NHTRDP 863 Grant No. 2009AA01Z329.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] E. Candes. Compressive sampling. *Int. Congress of Mathematics*, 3:1433–1452, 2006.

[3] D. Donoho. Compressive sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, Apr. 2006.

[4] D. B. Graham and N. M. Allinson. Characterizing virtual eigensignatures for general purpose face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, NATO ASI Series F, Computer and Systems Sciences, pages 446–456, 1998.

[5] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Proceedings of Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[6] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, 2007.

[7] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(5):684–698, 2005.

[8] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284, Nov. 1998.

[9] N. Ramanathan, R. Chellappa, and A. R. Chowdhury. Facial similarity across age, disguise, illumination and pose. In *Proceedings of the International Conference on Image Processing(ICIP)*, 2004.

[10] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 229(5500):2323–2326, Dec. 2000.

[11] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2), Feb. 2009.