EUSIPCO 2013 1569741443

GENERALIZED NON-LINEAR SPARSE CLASSIFIER

A. Majumdar, R. K. Ward and T. Aboulnasr

Department of Electrical and Computer Engineering, University of British Columbia
angshulm@ece.ubc.ca, rababw@ece.ubc.ca and taboulnasr@apsc.ubc.ca

ABSTRACT

In a recent study a novel classification algorithm called the Sparse Classifier (SC) assumes that if a test sample belongs to class k then it can be approximately represented by a linear combination of the training samples belonging to k. Good face recognition results were obtained by the SC method. This paper proposes two generalizations of the aforesaid assumption. The first generalization assumes that the test sample raised to a power can be approximated by a linear combination of the training samples of that class raised to the same powers. The second generalization assumes that the test samples raised to a power can be approximately represented by a non-linear combination of the training samples raised to the same power. The first generalization requires solving a group-sparse optimization problem with linear constraints while the second assumption requires solving a group-sparse optimization problem with non-linear constraints. We propose two greedy sub-optimal algorithms to solve the said problems. The classifiers developed in this work are used for single-image-per-person face recognition. We find that our first generalization leads to an improvement of 2-3% in recognition accuracy over SC, while the second generalization improves the recognition accuracy even further; about 6-7% better than the first generalization.

Index Terms— Greedy Algorithms, Classification, Face Recognition

1. INTRODUCTION

A recent work in face recognition [1] proposed a simple yet novel assumption: A test sample belonging to a particular class can be approximately expressed as a linear combination of the training samples of that class. This led to the Sparse Classification (SC) approach. With this simple assumption, very good recognition results were obtained on the Extended Yale and the AR face recognition databases [1].

The classification assumption of [1] is restrictive. We propose two generalizations. The first generalization assumption relaxes the condition that the test sample should be approximated by linear combination of the training samples of that class. We assume that the test samples raised to a certain power $p_i, i \in \{1,...,M\}$ can be approximated by a linear combination of the training samples raised to the same power. This is a generalization of the previous assumption, where only p=1 is considered. The first generalization assumption leads to a group sparse optimization problem with linear constraints. The second generalization assumption is more complex. It assumes that the test sample raised to a certain power $p_i, i \in \{1,...,M\}$ can be expressed approximately as a non-linear combination of the training samples raised to the same power. This also leads to a group sparse optimization problem but with non-linear constraints.

We are not aware of any work that addresses the problem of group-sparse optimization with non-linear constraints. Hence there is no algorithm to solve it efficiently. In this work, we propose an efficient greedy algorithm to solve the said problem.

In this work we are interested in the problem of face recognition when only a single training image of each person is available. More commonly it is referred to as the single-image-per-person recognition problem. A survey of this problem [2] shows that most of previous studies in this field employ the Nearest Neigbhour (NN) for classification. Studies like [3-9] differ from each other in their feature extraction method, but all use the same NN classification. Keeping the feature extraction the same, but by changing the classifier we will show that significant improvements in recognition results can be achieved.

The rest of the paper will be organized into several sections. Section 2, describes the Sparse Classification method [1]. In section 3, the proposed generalizations and the optimization tools needed to implement them are discussed. Section 4 discusses the experimental results. Finally in section 5, discussions and future scope of work are discussed.

2. LITERATURE REVIEW: SPARSE CLASSIFIER

The sparse classifier (SC) assumes that the training samples of a particular class approximately form a linear basis for a new test sample belonging to the same class. The assumption can be expressed formally as:

$$v_{k,test} = \alpha_{k,1} v_{k,1} + \alpha_{k,2} v_{k,2} + ... + \alpha_{k,n_k} v_{k,n_k} + \varepsilon \quad (1)$$

where $v_{k,i}$ are the training samples and ε is the approximation error.

Equation (1) expresses the assumption in terms of the training samples of a single (correct) class. Alternately, it can be expressed in terms of all the training samples in the form:

$$v_{k,test} = V\alpha + \varepsilon \tag{2}$$

where $V = [v_{1,1} | ... | v_{n,1} | ... | v_{k,1} | ... | v_{k,n_k} | ... v_{C,1} | ... | v_{C,n_C}]$

and $\alpha = [\alpha_{1,1} ... \alpha_{1,n_1} ... \alpha_{k,1} ... \alpha_{k,n_k} ... \alpha_{C,1} ... \alpha_{C,n_C}]'$.

According to the assumption, the solution to the inverse problem (2) should be sparse, i.e. only those coefficients in the vector α should be non-zeroes which correspond to the correct class of the test sample. The rest should all be zeroes. Ideally a sparse solution is achieved by solving the following optimization problem,

$$\min \| \alpha \|_0 \text{ such that } \| v_{k,test} - V\alpha \| < \sigma \tag{3}$$

However, solving this is an NP hard problem. For practical problems it can only be directly solved but only approximately by greedy algorithms or via convex approximations of the NP hard l_0-norm. There are many greedy and optimization based solvers to solve (3).

Once α is solved the classification proceeds as follows:

SC (Sparse Classifier) Algorithm [1]

1. Find a sparse solution to inverse problem (2).
2. For each class i repeat the following two steps:
   a. Find a representative sample for each class by a linear combination of the training samples belonging to that class by the equation $v_{rep}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j}$.
   b. Find the error between the reconstructed sample and the given test sample by $error(v_{test}, i) = \| v_{k,test} - v_{rep(i)} \|_2$.
3. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

3. PROPOSED GENERALIZATIONS

The full generalization over SC is achieved in two steps. In the first step, it is assumed that the test sample raised to certain powers can be approximately represented by a linear combination of training samples of the correct class raised to the same power (sub-section 3.1). In the second step, it is assumed that the test sample raised to certain powers can be approximated by a non-linear combination of the training samples of the correct class raised to the same power (sub-section 3.2).

3.1 GENERALIZED LINEAR SPARSE CLASSIFIER

The simple assumption in (1) says that a test sample can be approximately expressed as a linear combination of the training samples from the correct class. This is a simplistic assumption. We argue that this approximation may hold for several powers (p_1, …p_M) such that

$$\begin{aligned} v^{p_1}{}_{test} &= \alpha_{p_1,k,1} v^{p_1}{}_{k,1} + ... + \alpha_{p_1,k,n_i} v^{p_1}{}_{k,n_k} + \varepsilon_{p_1} \\ v^{p_2}{}_{test} &= \alpha_{p_2,k,1} v^{p_2}{}_{k,1} + ... + \alpha_{p_2,k,n_i} v^{p_2}{}_{k,n_k} + \varepsilon_{p_2} \\ &... \\ v^{p_M}{}_{test} &= \alpha_{p_M,k,1} v^{p_M}{}_{k,1} + ... + \alpha_{p_M,k,n_i} v^{p_M}{}_{k,n_k} + \varepsilon_{p_M} \end{aligned} \tag{4}$$

where v_p indicates that each coefficient of the sample v is raised to the power p.

We can write this expression (4) in terms of all the training samples of the class

$$v_{test}^{p_i} = V_{p_i}\alpha_{p_i} + \varepsilon_{p_i}, \forall i = 1:M \tag{5}$$

where V_{p_i} is a matrix formed by stacking the training samples raised to the power pi column-wise and ε_{p_i} is the error.

Thus (5) is a generalization of (2), where the latter forms a special case of the former where only a single power (p =) is considered.

To design a classifier based on our generalized assumption, the first task will be to organize the system of equations (4) and (5) in matrix-vector form:

$$\begin{bmatrix} v_{k,test}^{p_1} \\ v_{k,test}^{p_2} \\ ... \\ v_{k,test}^{p_M} \end{bmatrix} = \begin{bmatrix} V_{p_1} & 0 & ... & 0 \\ 0 & V_{p_2} & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & V_{p_M} \end{bmatrix} \begin{bmatrix} \alpha_{p_1} \\ \alpha_{p_2} \\ ... \\ \alpha_{p_M} \end{bmatrix} + \begin{bmatrix} \varepsilon_{p_1} \\ \varepsilon_{p_2} \\ ... \\ \varepsilon_{p_M} \end{bmatrix} \tag{6}$$

This can be expressed as

$$v = V\alpha + \varepsilon \tag{7}$$

where $v_{k,test} = [v_{k,test}^{p_1}, ..., v_{k,test}^{p_M}]'$;

$V = BlockDiag[V_{p_1}, ..., V_{p_M}]$ and $\alpha = [\alpha_{p_1}, ..., \alpha_{p_M}]'$

By definition (4) the structure of the coefficient vector α demands group sparsity, i.e. the indices in each of the α_p's should be non-zeroes for the correct class of the test sample. The groups are formed by the class of indices, i.e.

$$\alpha = [\underbrace{\alpha_{p_1,1}, \alpha_{p_2,1} ... \alpha_{p_M,1}}_{\alpha_1} ... \underbrace{\alpha_{p_1,C}, \alpha_{p_2,C} ... \alpha_{p_M,C}}_{\alpha_C}]'$$

where $\alpha_{p_j,i} = [\alpha_{p_j,i,1} ... \alpha_{p_j,i,n_i}]'$

With this notation, we frame the ideal group sparsity promoting optimization problem,

$$\min_{\alpha} \| \alpha \|_{2,0} \text{ such that } \| v_{test} - V\alpha \|_2 < \varepsilon \tag{8}$$

Solving the optimization problem (8) is NP hard. There are two approaches to solve it. The first one is to directly solve it using greedy sub-optimal algorithms. The second

one is to solve (8) by approximating the NP hard $l_{2,0}$-norm by a convex surrogate. The greedy algorithms are faster than convex optimization based algorithms. Therefore in this work we use a very fast algorithm called the Stagewise Block Orthogonal Matching Pursuit (StBOMP) [10] to recover a group sparse solution.

Once α is solved, we base our classification on a slight modification of the SC method:

GLSC (Generalized Linear Sparse Classifier) Algorithm

1. Solve the optimization problem expressed in (9) either by optimization or by greedy algorithm.
2. Find those i's for which $\|\alpha_i\|_2 > 0$. For those classes (i) satisfying the condition in step 2, repeat the following two steps:
    a. Obtain the representative a sample for each class by a linear combination of the training samples in that class via the equation $v_{rep}(i) = \sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j}$.
    b. Find the error between the reconstructed sample and the given test sample by $error(v_{test}, i) = \| v_{k,test} - v_{rep(i)} \|_2$
3. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

3.2 GENERALIZED NON-LINEAR SPARSE CLASSIFIER

In the second generalization step it is assumed that the test sample raised to certain powers can be approximately represented by a non-linear combination of the training samples raised to the same power, i.e. we are proposing an assumption of the form

$$v = f(V\alpha) + \varepsilon, \varepsilon \sim N(0, \sigma) \quad (9)$$

where

$v_{k,test} = [v_{k,test}^{p_1}, ..., v_{k,test}^{p_M}]'$; $V = BlockDiag[V_{p_1}, ..., V_{p_M}]$ and $\alpha = [\alpha_{p_1}, ..., \alpha_{p_M}]'$

This assumption opens a wide and powerful variety of possibilities in terms of modeling the classification problem since it breaks the restrictions imposed by linearity. The full generalized model comes at the cost of computational complexity. The final form of the classification assumption leads to an optimization problem of the form:

$$\min \| \alpha \|_{2,0} \text{ such that } \|v_{test} - f(V\alpha)\|_2 < \eta \quad (10)$$

This is an NP hard problem to solve. Recently a greedy algorithm for non-linear sparse system identification was proposed in [11]. It was meant for approximating sparse optimization problems of the form

$$\min \| \alpha \|_0 \text{ subject to } E(\alpha, v_{k,test}) \quad (11)$$

where $E(\alpha, v_{k,test})$ denotes an error measure not necessarily the l_2-norm (widely used for Gaussian Noise).

Our algorithm is based on ideas similar to [11] but is tailored for solving (14).

Greedy Non Linear Sparse Solution

Initialization - The sparse vector to be estimated is initialized to zero, $\alpha = 0$. The residual is initialized to the test sample, $r^{(0)} = v_{k,test}$. The set of chosen indices is empty $L^{(0)} = [\ ]$.

Iteration – Continue the following steps until the norm of the residual is less than a predefined value.

1. The first step computes the gradient of the error at the current coefficient estimate, i.e. $g = \frac{d}{d\alpha} \| v_{k,test} - f(V\alpha) \|_2^2 \ at\ \alpha^{(t-1)}$. This is basically a generalization of the OMP algorithm [22] where the correlations are the negative gradient of error term $\| v_{k,test} - V\alpha \|_2^2$ evaluated at the current coefficient estimate.
2. The group having index with highest gradient magnitude is chosen, $l = \{group(i) : \max | g(i)) | \}$. This step is also similar to OMP, where the index of the highest correlation is chosen.
3. The current set of indices is updated by adding the newly chosen indices $L^{(t)} = [L^{(t-1)}\ l]$.
4. The values of the signal at the chosen indices are computed by least squares optimization $x = \min \| v_{k,test} - f(V(:L^{(t)})x) \|_2$. This is a problem of non-linear least squares and does not have a closed form solution and needs to be solved iteratively.
5. The coefficient vector and the residual are updated, $\alpha(L^{(t)}) = x$ and $r^{(t)} = v_{k,test} - f(V\alpha)$.

This algorithm can be applied for a wide class of functions, the only restriction being $f(V\alpha) = 0, at\ \alpha = 0$.

The non-linear sparse estimation is the core behind the classification algorithm. Based on this estimation we propose the classification algorithm as follows:

GNSC (Generalized Non-linear Sparse Classifier) Algorithm

1. Solve the optimization problem expressed in (14) by the greedy algorithm.
2. Find those i's for which $\|\alpha_i\|_2 > 0$.
3. For those classes (i) satisfying the condition in step 2, repeat the following two steps:
    a. Obtain the representative a sample for each class by a linear combination of the training

samples in that class via the equation $v_{rep}(i) = f(\sum_{j=1}^{n_i} \alpha_{i,j} v_{i,j})$.

b. Find the error between the reconstructed sample and the given test sample by $error(v_{test}, i) = \| v_{k,test} - v_{rep(i)} \|_2$

4. Once the error for every class is obtained, choose the class having the minimum error as the class of the given test sample.

4. EXPERIMENTAL RESULTS

The proposed classification algorithms are applied on the problem of recognizing faces from a single training image of each person. We follow a similar experimental evaluation methodology as in [6]. Our evaluation is performed over the FERET database which consists of 14501 images of 1209 subjects. We only use the 3817 images (of 1200 subjects) that have the eye-position available, as we are interested only in face recognition and not face detection. The eye positions are required a priori for carrying forth the standard preprocessing steps from the FERET protocol.

Of the 1200 subjects, 226 subjects have 3 images per subject. In [6], it is suggested that this set be used as the generic gallery. These 678 images are also used for tuning our classifier. The training and testing datasets are formed from 1703 images which consist of at least 4 images per subject for another 256 persons. The training dataset is formed by randomly selecting 256 images (one image for each person) for the 256 people; the remaining 1447 images form the testing set.

The objective of this work is to show how the classification accuracy is improved by changing the classification from simple Nearest Neighbour (NN) classification to more sophisticated techniques like the SC and its proposed generalizations. Therefore we do not introduce any novelty into the feature extraction techniques. The feature selection methods used in this work are $(PC)^2A$ [4], SPCA [5], Eigenface Selection [6], SPCA+ [5] and sampled FLDA [8].

Owing to limitations in space we can not tabulate results for different number of Eigenfaces/Fisherfaces and use 40 feature points (Eigenfaces/Fisherfaces). In Table 1, we show how the classification accuracy improves for a fixed number of feature points when the classification algorithm gets progressively more sophisticated.

The NN and the SC are non-parametric classifiers. But certain parameters need to be decided for our proposed classifiers. For the Generalized Linear Sparse Classifiers (tables 3 and 4), we found that the values of index between 0.1 and 2 give good recognition accuracy. In this work we considered the values of p – 0.125, 0.25, 0.5, 1 and 2. We tried sampling the range uniformly (0.1 to 2 in steps of 0.1) but saw that there was no gain in recognition accuracy with such fine sampling.

The Generalized Non-linear Sparse Classifier offers a wide range of modeling functions to be used for classification. It is not possible to test all the different functional forms and decide the best one for our problem. In this work, we tested the following functions:

$f_1(A, x) = (Ax)^2 + Ax$

$f_2(A, x) = (Ax)^{1/2} + (Ax)^2 + Ax$

$f_3(A, x) = (Ax)^{1/2} + Ax$

Of these we found that the third function gives the best recognition results. The results are shown in table 5.

Table 1. Variation in Recognition Accuracy

Feature Extraction	NN	SC	GLSC	GNSC
$(PC)^2A$	0.48	0.5	0.54	**0.58**
SPCA	0.51	0.52	0.56	**0.59**
Eigenface Selection	0.54	0.55	0.57	**0.64**
SPCA+	0.52	0.56	0.58	**0.64**
sampled FLDA	0.5	0.51	0.54	**0.60**

The following points can be noted:

- The Sparse Classifier (SC) is always better than the Nearest Neighbour (2-3% improvement).
- The GLSC gives better results than the NN and the SC. It shows about 2-3% improvement over the SC.
- The GNSC gives considerably better results compared to the others. It shows about 6-7% improvement in recognition accuracy over the GLSC.

5. CONCLUSION

This work proposes major generalization of the sparse classification framework. The proposed classifiers were tested on the real-life problem of identifying faces of people from a single training image. The results show major improvement over previous Nearest Neighbour based methods.

This paper is exploratory in nature. The classification algorithms are highly generalized and flexible. But in order to make good use of these classifiers several questions must be answered – The first being the choice on the values of p for other classification problems (not necessarily face recognition). In our case, we found the values manually. The second question is even more important – how to choose the non-linear classification model. Again in this case, we tried several simple models and found the one that suits us the best.

REFERENCES

[1] Yang, Y., Wright, J., Ma, Y., Sastry, S. S., (to appear). Feature Selection in Face Recognition: A Sparse Representation Perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Tan, X., Chen, S., Zhou, Z. H, Zhang, F., 2006. Face recognition from a single image per person: A survey. Pattern Recognition 39 (9), 1725-1745.

[3] Wu, J., Zhou, Z. H., 2002. Face Recognition with one training image per person. Pattern Recognition Letters, 23(14), 1711-1719.

[4] Chen, S. C., Zhang D.Q., Zhou, Z. H., 2004. Enhanced (PC)2A for face recognition with one training image per person. Pattern Recognition Letters, 25(10), 1173-1181.

[5] Zhang D.Q., Chen S.C., Zhou, Z. H., 2005. A new face recognition method based on SVD perturbation for single example image per person. Applied Mathematics and Computation 163(2), 895-907.

[6] Wang J., Plataniotis K.N., Venetsanopoulos A. N., 2005. Selecting discriminant eigenfaces for face recognition. Pattern Recognition Letters 26(10), 1470-1482.

[7] Jung, H. C., Hwang, B. W., Lee, S. W., 2004. Authenticating Corrupted Face Image Based on Noise Model. International Conference on Automatic Face and Gesture Recognition, 272-277.

[8] Yin, H., Fu, P., Meng, S., 2006. Sampled FLDA for face recognition with single training image per person. Neurocomputing 69, 2443-2445.

[9] Majumdar, A., Ward, R. K., 2008. Single Image per Person Face Recognition with Images Synthesized by Non-Linear Approximation. International Conference on Image Processing, 2740-2743.

[10] Majumdar, A., Ward, R. K., 2009. Fast Group Sparse Classification. IEEE Canadian Journal of Electrical and Computer Engineering, 34 (4), 136-144

[11] T. Blumensath and M. E. Davies; "Gradient Pursuit for Non-Linear Sparse Signal Modelling", European Signal Processing Conference (EUSIPCO), Lausanne, Switzerland, April 2008.