Journal of AI and Data Mining
Vol 5, No 2, 2017, 223-234

Face Recognition using an Affine Sparse Coding approach

M. Nikpour[1], M.- R Karami[1*] and R. Ghaderi[2]

1. Electrical and Computer Engineering Department, Babol Noushirvani University of Technology, Babol, Iran.
2. Nuclear Engineering Department, Shahid Beheshti University of Tehran, Tehran, Iran.

Received 19 December 2016; Received 14 January 2017; Accepted 23 January 2017
*****Corresponding author: Mkarami@nit.ac.ir (M.Karami).*

Abstract
Sparse coding is an unsupervised method that learns a set of over-complete bases to represent the data such as image and video. Sparse coding has had an increasing attraction for image classification applications in the recent years. However in the cases where there are some similar images from different classes, such as face recognition applications, different images may be classified into the same class, and hence the classification performance may be decreased. In this paper, we propose an Affine Graph Regularized Sparse Coding approach for the face recognition problem. Experiments performed on several well-known face datasets show that the proposed method can significantly improve the face classification accuracy. In addition, some experiments are performed to illustrate the robustness of the proposed method to noise. The results obtained show the superiority of the proposed method in comparison to some other methods in face classification.

Keywords: *Sparse Coding, Manifold Learning, Face Recognition, Graph Regularization.*

1. Introduction

Face recognition is a significant task in image processing and computer vision studies. It is a challenging problem due to two reasons. Firstly, the face images of individual persons are mostly like each other and secondly, the face images are captured under challenging conditions like different poses, different conditions, and different illuminations [1].

Many methods have been introduced for face recognition in the recent years [2-5]. One of the appropriate methods used in this field is the sparse coding-based approach [6]. Sparse coding can represent images using a few active coefficients [7]. Accordingly, the interpretation and application of the sparse representations are easy, and simplify many image processing operations such as image classifications [8].

One of the most important targets in sparse coding applications is preserving the quality of sparse representation. In order to achieve this target, many works have been done to modify the sparsity constraint. In [9], the authors have added a nonnegative constraint to the objective function of the basis sparse coding method for improving the sparse coding method. In [3], the authors have analyzed the working mechanism for sparse representation-based classification, and have indicated that the collaborative representation sparsity makes this method powerful for face classification.

In [10], the authors have proposed a face recognition method based on the discriminative locality preserving vectors. In [11], the authors have improved the sparse coding method by adding a Laplacian term. In [12], the authors have proposed a sparse and dense hybrid representation (SDR) framework to alleviate the problems of sparse representation-based classification (SRC).

When the images are similar, the dictionary learned from the images cannot effectively encode the manifold structure of the images, and the similar images from different classes may be classified in the same class accordingly. Many research work have been done on dictionary learning. In [13], the authors have used the Fisher discrimination dictionary learning for sparse representation. In [14], the authors obtained a robust and reliable dictionary to improve the

performance of dictionary learning algorithms for face recognition. At first, the virtual face images are produced, and then an elaborate objective function is designed, and based on this objective function, they obtain an efficient algorithm to generate a robust dictionary.

Similar images lie on a manifold structure, and the images from different classes lie on different manifold structures [15]. It has been shown that if the geometrical structure is used and the local invariance is considered, the learning performance can be significantly improved. Recently, many literatures have focused on manifold learning problems, which represent the samples from different manifold structures. To preserve the geometrical information of the data, the authors in [16] have proposed to extract a good feature representation through which the manifold structure of data is spotted. The other methods such as graph regularization [11] and using weighted ℓ2-norm constraint are also introduced for improving the sparse representation. In [17], the authors have proposed a graph-based algorithm, called Graph regularized Sparse Coding (GraphSC), to give sparse representations that well-consider the local manifold structure of the data. Using Graph Laplacian as a smooth operator, the sparse representations obtained vary smoothly along the geodesics of the data manifold.

For solving the sparse coding problems, the authors in [15] have proposed a feature sign search method. This method reduces the non-differentiable problem to an unconstrained quadratic programming (QP). This problem can be solved rapidly by the optimization process. Our work also uses the feature sign search method to solve the proposed AGRSC optimization problem. For adapting the dictionary to achieve sparse representation, the authors in [18] have proposed a K-SVD method to learn the dictionary using orthogonal matching pursuit or basis pursuit.

Regarding the recent improvements in sparse coding and manifold learning, the two main problems of face recognition can still be investigated. We propose an Affine Graph Regularized Sparse Coding (AGRSC) algorithm to construct robust sparse representations for classifying similar images accurately. Specifically, the objective function of sparse coding has incorporated the Affinity term to make similar faces far from each other. Moreover, to improve the objective function with more discriminating power in data representation, we also incorporated the graph Laplacian term of coefficients in our objective function. This term can consider the geometrical structure of the data space by taking into account the local manifold structure of the data [17]. The experimental results verify the effectiveness of our AGRSC approach.

This paper is continued as follows: In section 2, the sparse coding and graph regularized sparse coding are described. In section 3 the proposed method is introduced. The experimental setup and results are indicated in section 4 and consequently, conclusions are drawn in section 5.

2. Preliminaries

This section introduces sparse coding and affine graph regularized sparse coding which are employed in this paper.

2.1. Sparse coding

Assume a data matrix Y = $[y_1, \ldots, y_n] \in R_{m\times n}$where n is the number of samples in the m-dimensional feature space. Let $\Phi = [\varphi_1, \ldots, \varphi_k] \in R_{m\times k}$ be the dictionary matrix where each column φ_i represents a basis vector in the dictionary, and X = $[x_1, \ldots, x_n] \in R_{k\times n}$ be the coding matrix where each column $\boldsymbol{x_i}$ is a sparse representation for a data point $\boldsymbol{y_i}$. Assuming the reconstruction error for a data point follows a zero-mean Gaussian distribution with isotropic covariance, while taking a Laplace prior to the coding coefficients and a uniform prior to the basis vectors, the maximum posterior estimate of Φ and X given Y is reduced to:

$$\min_{\Phi,X} \|Y-\Phi X\|_F^2 + \alpha \sum_{i=1}^{n} |x_i| \quad st. \|\varphi_j\|^2 \le c, \forall i=1,2,...,k \tag{1}$$

In the above equation α is a parameter for regularizing the level of sparsity of the codes obtained and the approximation of initial data. The objective function in (1) is not convex in Φ and X, and therefore, solving the above equation is not easy in this case. However, it is convex in either Φ or X. Therefore, solving this problem is done by alternatively optimizing Φ while fixing X and vice versa. As a result, the above-mentioned problem can be split into two reduced least squares problems: an $\ell 1$-regularized and an $\ell 2$-constrained, both of which can be solved efficiently by the existing optimization software [15].

2.2. Graph regularized sparse coding

In [17], the authors have proposed a method called the Graph Regularized Sparse Coding (GraphSC) method, which considers the manifold assumption to make the basis vectors with respect to the intrinsic geometric structure underlying the

input data. This method assumes that if two data points y_i and y_j are close in the intrinsic geometry of data distribution, then their codes φ_i and φ_j are also close. Consider a set of *n*-dimensional data points $y_1,...,y_n$;,GraphSC constructs a *p*-nearest neighbor graph *G* with *n* vertices each representing a data point. Let *W* be the weight matrix of *G*, if y_i is among the *p*-nearest neighbor of y_j, $W_{i,j}=1$; otherwise, $W_{i,j}=0$. $D=diag(d_1,...,d_n)$, $d_i=\sum_{j=1}^{n}W_{ij}$ and graph Laplacian $L=D-W$. A reasonable criterion for preserving the geometric structure in graph *G* is to minimize:

$$\frac{1}{2}\sum_{i,j=1}^{n}\left\|x_i-x_j\right\|^2 W_{i,j}=Tr(XLX^T). \quad (2)$$

By replacing the result into (1), GraphSC [1] is obtained:

$$\min_{\Phi,X}\left(\left\|Y-\Phi X\right\|_F^2+\gamma Tr(XLX^T)+\alpha\sum_{i=1}^{n}\left|x_i\right|\right). \quad (3)$$
$$st. \quad \left\|\varphi_i\right\|^2\le c,\forall i=1,2,...,k$$

In (3), γ is a parameter for regularizing the weight between sparsity of the codes obtained and preserving the geometrical structure.

3. Proposed method

In this section, we present the AGRSC algorithm for robust face recognition, which extends GraphSC by taking into account the affinity constraints on the samples. In the proposed method, at first, the Histogram of Gaussian (HOG) descriptor is extracted from face images. This descriptor is used in computer vision and image processing for the purpose of object detection. The technique counts the occurrences of gradient orientation in localized portions of an image. Due to the high dimensions of the descriptors, Principle Component Analysis (PCA) is applied to reduce dimensions of the descriptors. Then the sparse codes are extracted with the proposed method and at last these codes are classified with the Support Vector Machine (SVM) classifier. In figure 1, one can see the steps involved in the proposed method.

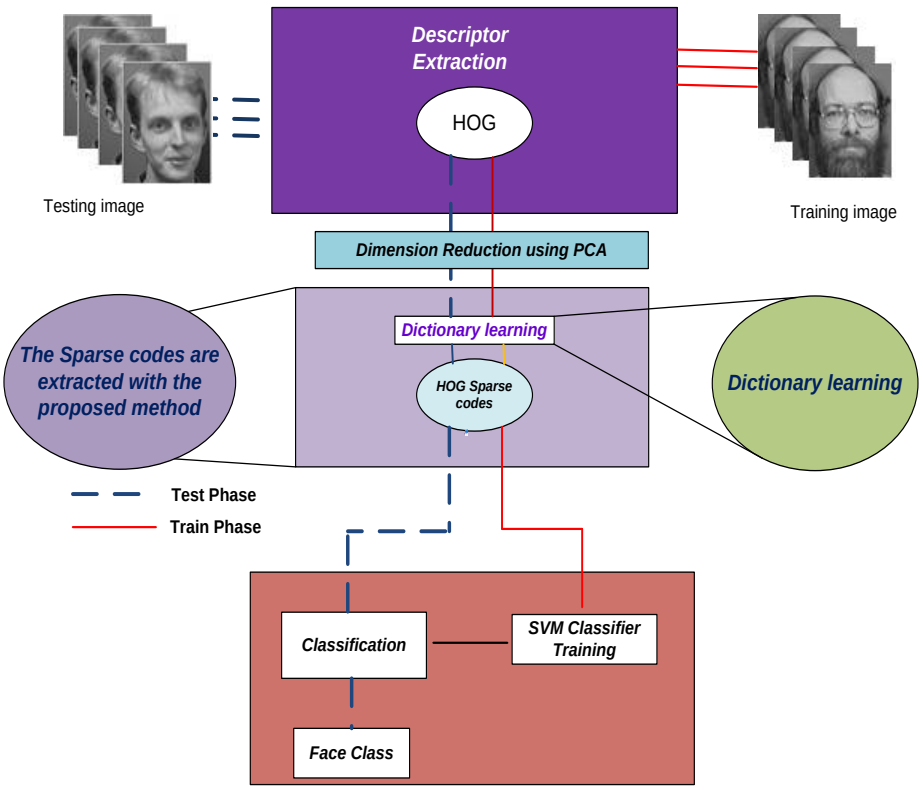


Figure 1. Diagram of proposed method for face recognition.

In the next section, before the AGRSC descriptions, we give some explanations for the HOG feature extraction.

3.2. Feature extraction

As mentioned earlier, the HOG feature is extracted for the sparse coding step. The HOG [19] characterizes the local object appearance and shape of faces by the local intensity gradients or edge direction distribution.

Assume that *P* is a facial image and *P(x,y)* is the intensity of pixel at the *(x,y)* coordinate.

The process of HOG extraction is shown in figure 2. At first, the HOG descriptor of the facial image is divided into some blocks and each block is subdivided into smaller squares called cells.

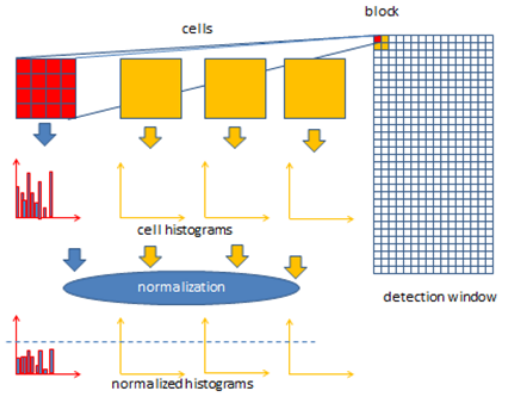


Figure 2. Extraction procedure of HOG feature

The histogram for each cell is computed using (4).

$$C_k = \sum_{k=1}^{9} V(x,y)$$
$$V(x,y) = \begin{cases} G(x,y) & \theta(x,y)\, in \quad bin_k \\ 0 & \theta(x,y)\, not \quad in \quad bin_k, \end{cases} \tag{4}$$

where, *G(x,y)* and $\theta(x,y)$ are the amplitude and directions of gradients at each pixel, respectively, and calculated using (12), and the gradient direction in the interval [0, π] is divided into 9 bins.

$$G(x,y) = \sqrt{G_x^2(x,y) + G_y^2(x,y)}$$
$$\theta(x,y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right). \tag{5}$$

where, $G_x(x,y)$ and $G_y(x,y)$ are the horizontal and vertical gradients respectively, computed using (6).

$$G_x(x,y) = P(x+1,y) - P(x-1,y)$$
$$G_y(x,y) = P(x,y+1) - P(x,y-1). \tag{6}$$

After the histogram extraction for each block, the histograms are normalized as follows:

$$N = \sqrt{\frac{C}{(\|C\|_1 + e)^2}}, \tag{7}$$

where, *e* is a small positive value in the case of an empty cell. At last, the histograms are combined to obtain the HOG feature representing the facial image.

3.3. Sparse code extraction

In linear sparse coding, a collection of k atoms $\varphi_1, \varphi_2, \ldots, \varphi_k$ is given that forms the columns of the over-complete dictionary matrix Φ. For extracting the sparse codes from the descriptors, we should have dictionary atoms for sparse code extraction, and this, in turn, needs a dictionary learning step.

The dictionary learning was done using the method proposed in [13]. We used a Fisher discrimination-based (FDDL) method. A structured dictionary $\Phi = [\varphi_1, \varphi_2, \ldots, \varphi_c]$ is learned instead of learning a shared dictionary to all classes, where φ_i is the class-specified sub-dictionary associated with class i, and c is the total number of classes. Assume that Y= $[y_1, y_2, \ldots, y_c]$ is the set of training samples, where y_i is the sub-set of the training samples from class I, X= $[x_1, x_2, \ldots, x_c]$ represents the coding coefficient matrix of Y over Φ, and x_i is the coding coefficients of y_i over Φ. The learning process uses the Fisher discrimination criterion [13]. Based on this criterion, the dictionary atoms are imposed on the coding coefficients so that they have small within-class scatter but big between-class scatter. This property could improve the facial image classification accuracy significantly.

The sparse codes of a feature vector $y \in R^m$; with a *l0*-minimization problem can be determined:

$$\min_{W \in R^m} \|W\|_0, s.t. \quad x = G_\Phi(W), \tag{8}$$

where, the function G_Φ is defined as $G_\Phi(w) = \Phi w$. In the proposed AGRSC method, the main technical difficulty is the proper interpretation of the function $G_\Phi(w)$ in the manifold setting, where the atoms $\varphi_1, \varphi_2, \ldots, \varphi_k$ are points in *M*, and Φ denotes the set of atoms, and due to the non-linearity property in this case, it is no longer possible to create a matrix with atoms. Moving to the more general manifold setting, we have forsaken the vector space structure in R^m.

In the linear sparse coding, each point is considered as a vector whose definition requires a reference point. However, in the AGRSC setting, each point cannot be considered as a vector and

therefore, must be considered as a point. This particular viewpoint is the main source of differences between the linear and AGRSC sparse codings.

Mathematically, we can view an image as a point in a high dimensional vector space whose dimensionality is equal to the number of pixels in the image [20]. Therefore, the descriptors extracted from facial images are on a high dimension manifold. When two facial images are similar to each other, these manifolds are overlapped in some sections. In order to solve this problem, in this paper, a new method is proposed to modify the usual notion of sparsity by adding an affine constraint to reduce the feature vector dimension on a manifold. A vector y is defined as an affine sparse vector if it can be written as follows [21]:

$$y = w_1\varphi_1 + w_1\varphi_1 + ... + w_n\varphi_n$$
$$w_1 + w_2 + ... + w_n = 1. \quad (9)$$

According to the definition, if the vector is constructed with combination of the affine samples, it can be mapped on the space with the lower dimension. The extracted sparse code vectors are in the space with high dimension manifold. Representing these vectors in places where the manifolds have interferences is very challenging. However if the facial images in a data set are effectively parameterized by a small number of continuous variables, then they will lie on or near a low-dimensional manifold in this high-dimensional space [22]. For representing a vector, if the sample selections are done based on only the nearest neighbors and the sparsity term, some of the samples may be selected from the irrelevant manifold; however, if the selected samples have the affinity constraint in addition, since the samples can be considered on the manifold with locally low dimension, only the samples on the relevant manifold could be selected. For a better perception of the proposed method, see figure 3.

Two overlapped manifolds are shown in the figure. Figure 3a indicates a representation of the samples a,b, and c regarding only the sparsity term, and figure 3b indicates the representation of the same points regarding the manifold constraints in addition to the sparsity constraint. The samples A and B in both figures 3a and 3b are represented by the atoms from the corresponding manifolds correctly. These two samples haven't any conflict with the other manifold.

Sample c is under different conditions. As indicated earlier, this sample is located on the green manifold. If you represent this sample with its adjacent atoms, and only consider the sparsity term, we should consider the other manifold samples for representation the same as figure 3.a. However, if we consider the GraphSc ($Tr(XLX^T)$) and Affinity terms for its representation in addition, we will reach a better conclusion. As previously pointed out, the term $Tr(XLX^T)$ emphasizes on the problem that if the samples of a manifold are close to each other, their codes will be close to each other as well. Also the Affinity constraint forces a collection of the closest neighbors of the concerned dictionary atoms for representing every sample. Therefore, a collection of weights for every sample are chosen in a way that every point is represented by a linear combination of its neighbors. The former samples are located on a manifold with high dimensions and the objective of the Affinity term is to reduce its dimensions. The characteristic of this new term causes sample c to be represented with utilization of the concerned manifold data (Figure 3b).

According to the above-mentioned descriptions, we can add an affinity term to (1):

$$\min_{\Phi,X}\left(\left\|Y-\Phi X\right\|_F^2+\gamma Tr(XLX^T)+\alpha\sum_{i=1}^{n}\left|x_i\right|\right) \quad (10)$$
$$st.\ \ \sum_{i=1}^{n}x_i=1.$$

The constraint term $\sum_{i=1}^{n} x_i = 1$ is added to the main criterion as a lagrangian coefficient, which leads to:

$$\min_{\Phi,X}\left\|Y-\Phi X\right\|_F^2+\gamma Tr(XLX^T)+\alpha\sum_{i=1}^{n}\left|x_i\right|+\beta(1-\sum_{i=1}^{n}x_i)^2, \quad (11)$$

where, β is a parameter for tuning the affinity constraint. To tune parameters $\alpha, \beta,$ and γ, some experiments were done, which are described in the next section.

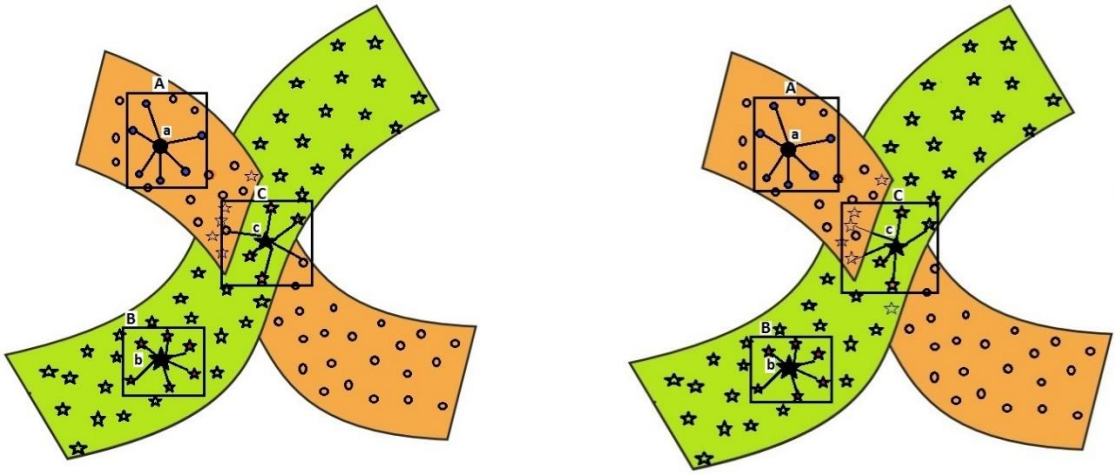
3.4. Solution of AGRSC

We applied the feature-sign search algorithm [15] to solve the optimization problem (11).

To solve non-differentiable problems in non-smooth optimization methods, a necessary condition for a parameter vector to be a local minimum is that the zero-vector should be a member of the sub-differential set containing all the sub-gradients in the parameter vector [23].

Following [17, 23], the optimization of AGRSC was divided into two steps: 1) $\ell 1$-regularized least squares problem; the affine graph regularized sparse codes X were learned with dictionary Φ fixed, and 2) $\ell 2$-constrained least squares problem; the dictionary Φ was learned with affine

graph regularized sparse codes *X* fixed. The above two steps were repeated, respectively, until a stop criterion was indulged.



a) Representation of samples a, b, and c without Affinity constraint.

b) Representation of samples a, b, and c with Affinity constraint.

Figure 3. Effectiveness of Affinity constraint in representation of samples from overlapped manifolds.

The optimization problem in the first step can be solved by optimizing over each x_i individually.

Since (11) with *l1*-regularization is non-differentiable when x_i contains the value of 0, to solve this problem, the standard unconstrained optimization methods cannot be applied. Several approaches have been proposed to solve the problem of this form [11]. In what follows, we introduce an optimization method based upon coordinate descent to solve this problem [24]. It can easily be seen that (11) is convex, thus the global minimum can be achieved.

We updated each vector individually by holding all the other vectors constant. In order to solve the problem by optimizing over each x_i, we should re write (11) in a vector form. The reconstruction error $\|Y-\Phi X\|_F^2$ can be re written as:

$$\sum_{i=1}^{m}\|Y-\Phi X\|_F^2 \qquad (12)$$

.

The Laplacian regularizer $Tr(XLX^T)$ can be rewritten as:

$$Tr(XLX^T)=Tr\left(\sum_{i,j=1}^{n}L_{i,j}x_i x_j^T\right)= \sum_{i,j=1}^{n}L_{ij}x_i x_j^T=\sum_{i,j=1}^{n}L_{ij}x_i^T x_j. \qquad (13)$$

Combining (11), (12), and (13), the problem can be written as:

$$\min\sum_{i=1}^{n}\|y_i-\Phi x_i\|_F^2+\gamma\sum_{i,j=1}^{n}L_{ij}x_i^T x_j+ \alpha\sum_{i=1}^{n}|x_i|+\beta\left(1-\sum_{i=1}^{n}x_i\right)^2. \qquad (14)$$

When updating x_i, the other vectors $x_{j\ i\neq j}$ are fixed. Thus we get the following optimization problem:

$$\min_{x_i}G(x_i)=\min\begin{pmatrix}\|y_i-\Phi X\|^2+\gamma L_{ii}x_i^T x_i+x_i^T H_i+\\ \alpha\sum_{j=1}^{k}|x_i^{(j)}|+\beta\left(1-\sum_{i=1}^{n}x_i\right)^2\end{pmatrix} \qquad (15)$$

where $H_i=2\gamma\left(\sum_{j\neq i}L_{ij}s_j\right)$ and $x_i^{(j)}$ is the j-th coefficient of x_i.

Following the feature-sign search algorithm proposed in [25]), (15) can be solved as follows. In order to solve the non-differentiable problem, we adopt a sub-gradient strategy, which uses sub-gradients of $G(x_i)$ at non-differentiable points. Primarily we define:

$$p(x_i)=\|y_i-\Phi x_i\|^2+\gamma L_{ii}x_i^T x_j+ x_i^T H_i+\beta\left(1-\sum_{i=1}^{n}x_i\right)^2. \qquad (16)$$

Then,

$$G(x_i)=p(x_i)+\alpha\sum_{j=1}^{k}|x_i^{(j)}|. \qquad (17)$$

Recall that a necessary condition for a parameter vector to be a local minimum in non-smooth optimizations, is that in the set containing all sub-gradients at this parameter vector, the zero-vector is an element of the sub-differential [23]. We define $\nabla_i^{(j)}|x_i|$ as the sub-differentiable value of the jth coefficient of x_i. If $|x_i^{(j)}| > 0$, then the absolute value function $|x_i^{(j)}|$ is differentiable, and therefore, $\nabla_i^{(j)}|x_i|$ is given by $sign(x_i^{(j)})$. If $x_i^{(j)} = 0$, then the subdifferentiable value $\nabla_i^{(j)}|x_i|$ is set [-1,1]. Thus the optimality condition for achieving the optimal value for $G(x_i)$ is:

$$\begin{cases} \nabla_i^{(j)} p(x_i) + \alpha sign(x_i^{(j)}) & if \quad |x_i^{(j)}| > 0 \\ \left|\nabla_i^{(j)} p \;\; x_i\right| \le \alpha & if \quad x_i^{(j)} = 0 \end{cases} \quad (18)$$

Then we consider how to select the optimal sub-gradient $\nabla_i^j G(x_i)$, when the optimality conditions are violated, i.e.; in the case that $\left|\nabla_i^{(j)} p \;\; x_i\right| > \alpha$ if $x_i^{(j)} = 0$. When $x_i^{(j)} = 0$, we consider the first term in the previous expression $\nabla_i^{(j)} p(x_i)$. Suppose that $\nabla_i^{(j)} p(x_i) > \alpha$; this means that $\nabla_i^{(j)} G(x_i) > 0$, regardless of the sign of $x_i^{(j)}$. In this case, in order to decrease $G(x_i)$, we should decrease $x_i^{(j)}$. Since $x_i^{(j)}$ starts at zero, the very first infinitesimal adjustment to $x_i^{(j)}$ will make it negative. Therefore, we can let $sign(x_i^{(j)}) = -1$. Similarly, if $\nabla_i^{(j)} p(x_i) < -\alpha$, then we let $sign(x_i^{(j)}) = 1$. To update x_i, suppose that we know the signs of $x_i^{(j)}{}'s$ at the optimal value; then we can remove the $l1$-norm on $x_i^{(j)}$ by replacing each term $|x_i^{(j)}|$ with either $x_i^{(j)}$ (if $x_i^{(j)} > 0$) or $-x_i^{(j)}$ (if $x_i^{(j)} < 0$) or 0 (if $x_i^{(j)} = 0$). Thus; (13) is converted to a standard unconstrained QP. In this case, the problem can be solved by a linear system. The algorithmic procedure of learning affine graph regularized sparse codes can be described as follows:

- for each x_i, search for signs of $sign(x_i^{(j)}); i = 1,\dots,k$
- solve the reduced QP problem to get the optimal x_i^* that minimizes the objective function
- return the optimal coefficients matrix $X^* = [x_1^*, x_2^*, \dots, x_n^*]$

In the algorithm, we maintain an active set $A = \{ j \mid x_i^{(j)} = 0, \left|\nabla_i^{(j)} p(x_i)\right| > \alpha \}$ for potentially non-zero coefficients and their corresponding signs $\theta = [\theta_1, \dots, \theta_k]$, while updating each x_i. Then it systematically searches for the optimal active set and coefficient signs that minimize the objective function (9). In each activating step, the algorithm uses the zero-value whose violation of the optimality condition $\nabla_i^{(j)} p(x_i) > \alpha$ is the largest.

The detailed algorithmic procedure of learning affine graph regularized sparse codes is stated in Algorithm 1.

Algorithm1: Learning affine graph regularized sparse codes

Input: $Y = [y_1, \dots, y_n]$, Φ, L, α, β, γ.

1- $1 \le i \le n$

2- **Initializing**: $x_i = \vec{0}, \theta = \vec{0}, A = \emptyset, \theta_j \epsilon \{-1,0,1\} = \text{sign}(x_i^{(j)})$.

3- **Activating:** $j = \arg\max_j \left|\nabla_i^{(j)} p(x_i)\right|$:

$if \quad \nabla_i^{(j)} p(x_i) > \alpha, \; \theta_j = -1, A = \{j\} \cup A$

$if \quad \nabla_i^{(j)} p(x_i) < -\alpha, \; \theta_j = 1, A = \{j\} \cup A$

4- Feature sign: $\hat{\Phi}$ is submatrix of Φ contains only columns corresponding to A . $\hat{x}_i$, $\hat{p}_i$ are subvectors of x_i , p.

$$\min u(\hat{x}_i) = \left\|y_i - \hat{\Phi}\hat{x}_i\right\|^2 + \gamma L_{ii}\hat{x}_i^T\hat{x}_i + \hat{x}_i^T\hat{H}_i + \beta(1 - \sum_{i=1}^{n} \hat{x}_i)^2 + \alpha\hat{\theta}^T\hat{x}_i^T$$

Let $(\partial u(\hat{x}_i)/\partial \hat{x}_i) = 0$

$$-2\hat{\Phi}^T(y_i - \hat{\Phi}\hat{x}_i) + 2\gamma\, L_{ii}\hat{x}_i + 2\gamma\left(\sum_{j \ne i} L_{ij}\hat{x}_i\right) + 2\beta(1 - 1^T\hat{x}_i)1 + \alpha\hat{\theta} = 0$$

$$\hat{x}_i^{new} = \left(\hat{\Phi}^T\hat{\Phi} + \gamma\, L_{ii}I + \beta 11^T\right)^{-1}\left(\hat{\Phi}^T y_i + \beta 1 - \frac{1}{2}(\alpha\hat{\theta} + \hat{H}_i)\right),$$

I is the identity matrix.

5- The optimality conditions:

Condition(1):

If $\nabla_i^{(j)} p(x_i) + \alpha sign\left(x_i^{(j)}\right) = 0, \forall \; x_i^{(j)} \ne 0$

Go to condition(2).

Else go to step 4

Condition(2):

If $\left|\nabla_i^{(j)} p(x_i)\right| \le \alpha, \forall\, x_i^{(j)} = 0$

Return x_i .

Else go to step 3

6- End

4. Experiments

In this section, for evaluating the proposed AGRSC approach, some experiments for image classification were performed. Some experiments were done on five well-known datasets including

ORL, Extended Yale B, FERET, AR, and LFW. These datasets contain several face images from distinct persons and under different conditions such as times, lighting, facial expressions and occlusions. Also some experiments were done on some noisy images with different variances for evaluating the robustness of the proposed method to noise.

4.2. Data preparation

ORL, Extended Yale B, FERET, AR, and LFW face databases are well-known datasets widely used in computer vision and pattern recognition research works. The experiments were done on these datasets. In continuation, we introduced these datasets.

Extended YaleB database [27]**.** This database contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions.

The images in the database were captured using a purpose-built illumination rig. This rig was fitted with 64 computer controlled strobes. The 64 images of a subject in a particular pose were acquired at camera frame rate (30 frames/second) in about 2 seconds, so there was only a small change in head pose and facial expression for those 64 (+1 ambient) images. The image with ambient illumination was captured without a strobe going off.

FERET database. The data is obtained from the UCI database [28]. It contains face images about 72 persons, and every body has 6 variations in expression.

AR database. This database consists of 126 subjects of over 4000 frontal face images [29]. These images have different illumination variations, facial expression and occlusion.

Following the standard evaluation procedure, we used a subset of the database consisting of 2600 images from 50 male subjects and 50 female subjects. For each person, we randomly selected 20 images for training and the other 6 for testing.

LFW dataset. The LFW dataset [30] contains 13233 images of 5749 individuals. The facial images in this dataset were taken in unconstrained environments. In figure 4, some examples of the datasets are shown.

4.3. Experimental setup

To evaluate the proposed AGRSC approach, the results of this method on five defined datasets were compared with some recent approaches; including LRC [2], CRC [31], SRC [8], LLC [4], FDDL [13], LH-ESRC [19], SPN-DL [14], D-KSVD [9], SVGDL [32], LC-KSVD [7], PDPL [33], and DNFC [34].

Following [17], AGRSC was performed on HOG descriptors extracted from the facial images. Before sparse code extraction from face data using the proposed method, the data dimensions were reduced using principle component analysis by keeping 98% of the information in the largest Eigen vectors. After applying the proposed algorithm to the reduced data, the SVM classifier with χ^2 kernel was applied on the sparse codes.

Under our experimental setup, we tuned the optimal parameters for the target classifier using the leave one subject out cross validation method. Therefore, we evaluated the proposed method on datasets by empirically searching the parameter space for the optimal parameter settings, and reported the best results.

Figure 4. Some samples of datasets. From top to bottom: ORL, Extended YALE B, FERET, AR And LFW datasets.

For the proposed AGRSC method, we set the trade-off parameters α, β, γ through searching. The parameter values using the ORL face dataset is shown in figure 5. As it can be seen, the parameters α, β, and γ were set to 30, 0.1 and 0.6, respectively.

At first, the value for the γ parameter was achieved for the best recognition rate, assuming $\alpha, \beta = 1$. As it can be seen in figure 5a, the highest recognition rate was achieved for $\gamma = 0.6$. In the next step, the value for α was achieved, assuming $\gamma = 0.6$ and $\beta = 1$ for the best recognition rate. As it can be seen in figure 5b, the best value for this parameter can be a number between 28 and 45. We set $\alpha = 30$, and using the same experiments, the best value for β was achieved to be 0.1.

As mentioned earlier, the dictionary learning process has been done using the method proposed in [13]. It should be noted that, the affinity constraint can be more successful when the sparsity is large enough because with the coefficients not enough sparsity, the coefficients

may be selected from the hyperplane with higher dimensions than data's original dimension. In this case, if the affinity constraint is added to the objective function, it can even worsen the performance with respect to the GraphSC method.

4.4. Experimental results

In order to evaluate the proposed method, it was performed on five well-known datasets.

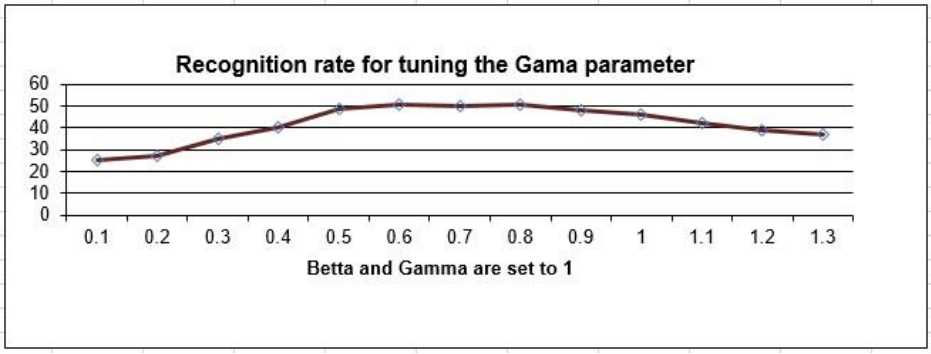
The classification accuracy of AGRSC on ORL dataset is illustrated in table 1 as a confusion matrix. As mentioned earlier, the ORL dataset contains 40 classes of faces. Due to the lack of space in the table, only 10 classes were depicted.

Among the whole dataset, classes 4 and 6, classes 8 and 10, classes 14 and 17, and classes 5 and 18 are very similar to each other. Therefore, we used these classes in addition to classes 1 and 2 in the confusion matrix to show the superiority of the proposed method in classifying face datasets in table 1. The mean recognition rate was 92%. When the other 30 classes were considered as well, the mean accuracy rate was raised up to 97.2%.

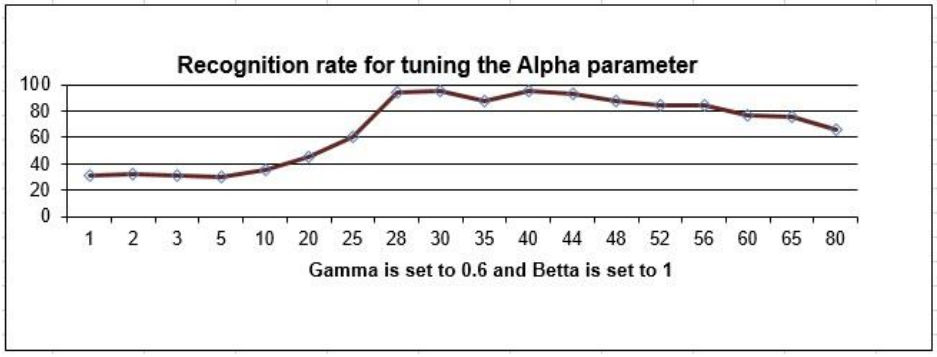
4.5. Robustness to noise

For a better evaluation of the proposed method, we aimed to test the robustness of our method at the presence of noise. Some Gaussian noise added to the database images and the experiments for evaluation are repeated for noise variance levels as 10, 20, 30, 40, and 50.

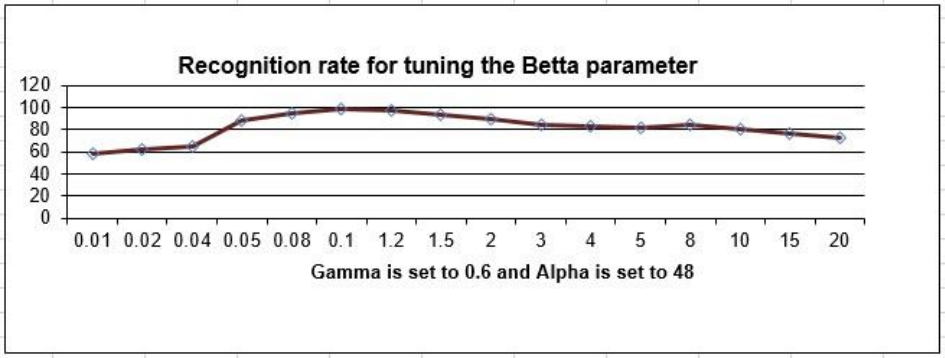
Figure 7 shows some test images under different levels of noise.



a) Recognition rate variations for gamma changes by setting Alpha=Beta=1.



b) Recognition rate variations for alpha changes by setting Gamma=0.6 and Beta=1.



c) Recognition rate variations for Beta changes by setting Gamma=0.6 and Alpha=30.

Figure 5. The parameters setting using ORL dataset.

Table 1. The confusion matrix for the proposed method on the ORL data

	C1	C2	C4	C5	C6	C8	C10	C14	C17	C18
C1	99	0	1	0	0	0	0	0	0	0
C2	0	100	0	0	0	0	0	0	0	0
C4	1	0	97	1	1	0	0	0	0	0
C5	0	0	4	87	0	3	0	1	5	0
C6	0	0	2	1	95	0	1	0	0	1
C8	0	0	0	0	0	95	4	1	0	0
C10	0	0	0	0	0	6	87	2	4	1
C14	0	0	0	0	0	0	1	90	4	5
C17	0	0	0	1	2	1	5	11	78	2
C18	0	0	0	1	1	0	2	4	0	92

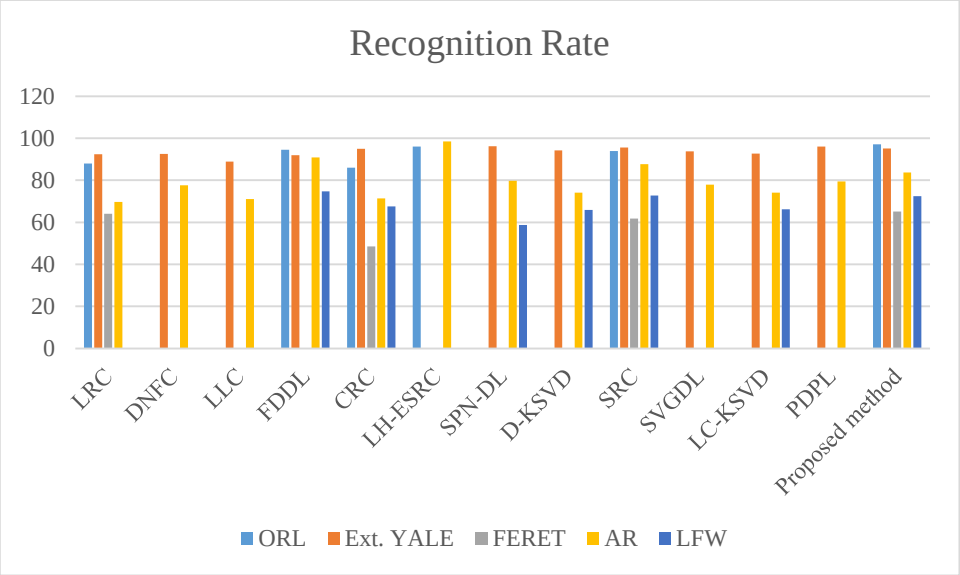


Figure 6. Recognition rate for proposed method in comparison with several other methods.

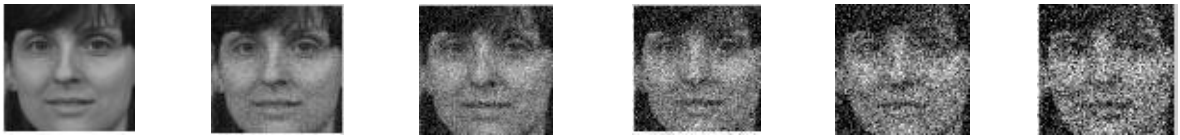


Figure 7. One test image under different levels of Gaussian noise. From left to right original image, images cluttered with Gaussian noise with variances 10, 20, 30, 40, and 50, respectively.

Table 2 shows the recognition rate of the proposed method under different noise levels for each database. It can be seen that the recognition rate of the proposed method for the ORL, Extended YALE B, FERET, AR, and LFW datasets are reduced only 3.8%, 3.24%, 3.53%, 4.02%, and 4.3%, respectively. This means that the proposed method is stable with noise.

5. Conclusion

In this paper, a novel approach was proposed for robust face recognition. In the proposed method, after extracting the HOG descriptors from the original face images, the sparse codes were extracted from the descriptors. For this purpose, the well-defined graph regularized sparse coding method was improved by adding the affinity constraint.

Using this term, until the sparsity was big enough, the manifold structure of features was better preserved. Finally, the codes obtained were classified with the SVM classifier. The results obtained indicated that the proposed AGRSC method in comparison with many other approaches had a better performance. The proposed method is efficient for face recognition for two reasons. Firstly, the dictionary atoms, because of the property of the FDDL method, has

enough discriminant, and secondly, the sparse codes extracted from the descriptors, because of the affinity characteristic, can more easily choose the correct class.

Table 2. Recognition rate for noisy images.

Database	Noise Variance	Recognition Rate
ORL	10	96.8
	20	95.4
	30	94.1
	40	92.4
	50	90.3
Extended Yale	10	94.7
	20	93.6
	30	92.1
	40	90.9
	50	88.5
FERET	10	64.6
	20	63.5
	30	62.3
	40	60.1
	50	57.3
AR	10	82.8
	20	81.3
	30	80.1
	40	78.5
	50	75.4
LFW	10	71.9
	20	70.7
	30	68.1
	40	66.1
	50	63.2

References

[1] Cao, F. Hu, H. Lu, J. Zhao, J. Zhou, Z. & Wu, J. (2016). Pose and Illumination Variable Face Recognition via Sparse Representation and Illumination Dictionary, Knowledge-Based Systems, doi: 10.1016/j.knosys.2016.06.001.

[2] Naseem, I. Togneri, R. & Bennamoun, M. (2010). Linear regression for face recognition, IEEE Trans. Neural Netw. Learn. Syst. vol. 32, no. 11, pp. 2106–2112.

[3] Zhang, L. Yang, M. & Feng, X. (2011). Sparse representation or collaborative representation: which helps face recognition? in: Proceedings of IEEE International Conference on Computer Vision, pp. 471–478.

[4] Wang, J. Yang, J. Yu, K. & Lv, F. (2010). Locality-constrained linear coding for image classification, in: Proceedings of IEEE International Conference on Computer Vision, pp. 3360–3367.

[5] Shafeipour, S. Seyedarabi, H. & Aghagolzadeh, A. (2016). Video-based face recognition in color space by graph-based discriminant analysis, Journal of AI and Data Mining, vol.4, no.2, pp. 193-201.

[6] Chen, W. Zhao, Y. Pan, B. & Chen, B. (2016). Supervised Kernel Nonnegative Matrix Factorization for Face Recognition Neurocomputing, http://dx.doi. org/ 10.1016/ j.neucom.2016.04.014.

[7] Jiang, Z. Lin, Z. & Davis, L. S. (2013). Label consistent K-SVD: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 11, pp. 2651–2664.

[8] Wright, J. Yang, A. Y. Ganesh, A. Sastry, S. S. & Ma, Y. (2009). Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no. 2, pp. 210-227.

[9] Liu, Y. N. Wu, F. Zhang, Z. H. Zhuang, Y. T. & Yan, S. C. (2010). Sparse representation using nonnegative curds and whey. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[10] Wen, Y. Zhang, L. von Deneen, K. M. & He, L. (2015). Face recognition using discriminative locality preserving vectors, Digit. Signal Process, http://dx.doi.org/10.1016/j.dsp.2015.11.001.

[11] Gao, S. I. Tsang, W.-H. Chia, L.-T. & Zhao, P. (2010). Local features are not lonely – laplacian sparse coding for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[12] Jiang, X. & Lai, J. (2014). Sparse and Dense Hybrid Representation via Dictionary Decomposition for Face Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence. ,DOI 10.1109/TPAMI.2014.2359453.

[13] Yang, M. Zhang, L. Feng, X. & Zhang, D. (2011). Fisher discrimination dictionary learning for sparse representation, In: IEEE International Conference on Computer Vision, pp. 543-550.

[14] Xu, Y. Li, Z. Zhang, B. Yang, J. & You, J. (2017). Sample diversity, representation effectiveness and robust dictionary learning for face recognition, Information Sciences, vol. 375, pp. 171–182.

[15] Lee, H. Battle, A. Raina, R. & Ng, A. Y. (2006). Efficient sparse coding algorithms. In Advances in Neural Information Processing Systems 20, NIPS.

[16] Zheng, M. Bu, J. Chen, C. Wang, C. Zhang, L. Qiu, G. & Cai, D. (2007). Graph Regularized Sparse Coding for Image Representation, Journal Latex Class Files, vol. 6, no. 1.

[17] Zheng, M. Bu, J. Chen, C. Wang, C. Zhang, L. Qiu, G. & Cai, D. (2011). Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing, vol. 20, no.5, pp. 1327-1336.

[18] Quanz, B. Huan, J. & Mishra, M. (2012). Knowledge transfer with low-quality data: A feature extraction issue. IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 10, pp. 1789 – 1802.

[19] Zheng, C. Hou, Y. & Zhang, J. (2015). Improved Sparse Representation with Low-Rank Representation for Robust Face Recognition, Neurocomputing, http://dx.doi.org/ 10.1016/ j.neucom. 2015.07.146.

[20] Turk, M. & Pentland, A. (1991). Eigenfaces for recognition, Jornal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86.

[21] Michiel, H. (2001). Affine transformation, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4.

[22] Lu, H. Fainman, Y. & Hecht-Nielsen, R. (1998). Image manifolds, Applications of Artificial Neural Networks in Image Processing III, Proceedings of SPIE, vol. 3307, pp. 52–63.

[23] Fletcher, R. (1987). Practical methods of optimization, Wiley-Interscience.

[24] Lu, X. Yuan, Y. & Yan, P. (2014). Alternatively Constrained Dictionary Learning for Image Superresolution, IEEE Transactions on Cybernetics, vol. 44, no. 3, pp.366-377.

[25] Cand`es, E. & Tao, T. (2006). Near-optimal signal recovery from random projections: niversal encoding strategies?, IEEE transactions on information theory, vol. 52, no. 12, pp. 5406–5425.

[26] Samaria, F. & Harter, A. (1994). Parameterisation of a Stochastic Model for Human Face Identification, Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL.

[27] Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Patt. Anal. Mach. Intell., vol. 23, no. 6, pp. 643-660.

[28] The Facial Recognition Technology (FERET) Database,http://www.itl.nist.gov/iad/humanid/feret/feret_master.html (last visited January 2017).

[29] Martinez, A. M. (1998). The AR face database, CVC Technical Report.

[30] Huang, G. B. Ramesh, M. Berg, T. & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments, University of Massachusetts, Amherst, Tech. Rep.

[31] Zhang, Q. & Li, B. (2010). Discriminative K-SVD for dictionary learning in face recognition, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2691–2698.

[32] Cai, S. Zuo, W. Zhang, L. Feng, X. & Wang, P. (2014). Support vector guided dictionary learning, in: Proceedings of European Conference Computer Vision, pp. 624–639.

[33] Gu, S., Zhang, L., Zuo, W. & Feng, X. (2014). Projective dictionary pair learning for pattern classification, in: Proceedings of Advances in Neural Information Processing Systems, pp. 793–801.

[34] Xu, Y. Fang, X. Li, X. Yang, J. You, J. Liu, H. & Teng, S. (2014). Data Uncertainty in Face Recognition, IEEE Trans. Cybern., vol. 44, no. 10, pp. 1950–1961.

نشریه هوش مصنوعی و داده کاوی

بازشناسی چهره با استفاده از رویکرد کدگذاری تنک افین

محسن نیک پور[1]، محمدرضا کرمی*[1] و رضا قادری[2]

[1] دانشکده برق و کامپیوتر، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران.

[2] دانشکده مهندسی هسته ای، دانشگاه شهید بهشتی تهران، تهران، ایران.

ارسال ۲۰۱۶/۱۲/۱۹؛ بازنگری ۲۰۱۷/۰۱/۱۴؛ پذیرش ۲۰۱۷/۰۱/۲۳

چکیده:

کدگذاری تنک یک روش بدون مربی است که مجموعه‌ای از پایه‌های فوق کامل را برای بازنمایی داده‌هایی مانند تصویر و ویدئو بکار می‌گیرد. در سال-های اخیر، توجه ویژه‌ای به روش کدگذاری تنک در کاربردهای کلاسبندی تصویر شده است. در حالی که در مواردی مانند کلاسبندی چهره که چند تصویر مشابه از کلاس‌های مختلف وجود دارد، تصاویر مختلف ممکن است در یک کلاس کلاسبندی شوند و در نتیجه کارایی کلاسبندی کاهش یابد. در این مقاله یک رویکرد کدگذاری تنک گراف منظم افین برای بازشناسی چهره ارائه شده است. آزمایشات روی چندین دادگان مشهور چهره نشان داده‌اند که روش پیشنهادی می‌تواند به خوبی دقت کلاسبندی چهره را بالا ببرد. همچنین آزمایشاتی نیز جهت ارزیابی قدرت روش پیشنهادی در شرایط نویزی نیز انجام شده‌اند. نتایج به دست آمده نشان‌دهنده برتری روش پیشنهادی نسبت به برخی از روش‌های دیگر در کلاسبندی چهره است.

کلمات کلیدی: کدگذاری تنک، یادگیری منیفلدی، بازشناسی چهره، تنظیم گراف.