Neurocomputing 140 (2014) 77–83

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A novel supervised feature extraction and classification fusion algorithm for land cover recognition of the off-land scenario

Yan Cui *, Zhong Jin, Jielin Jiang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form
24 December 2013
Accepted 1 March 2014
Communicated by M. Wang
Available online 8 April 2014

Keywords:
Feature extraction
Sparse representation
Dictionary learning
Land cover recognition
Classification

ABSTRACT

In this paper, a novel supervised feature extraction and classification fusion algorithm based on neighborhood preserving embedding (NPE) and sparse representation is proposed. Specifically, an optimal dictionary is adaptively learned to bate the trivial information of the original training data; then, in order to obtain the sparse representation coefficients, a sparse preserving embedding map is sought to reduce the dimensionality of high-dimensional data, and the test data is classified by the corresponding sparse representation coefficients. Finally, the novel supervised fusion algorithm is applied to the land cover recognition of the off-land scenario. Experimental results show that the proposed method leads to promising results in fusing feature extraction and classification.

1. Introduction

Techniques for dimensionality reduction in unsupervised and supervised learning tasks have attracted much attention in computer vision, machine learning and biometrics. Among them, subspace learning and manifold learning methods have been dominantly and successfully used in dimensionality reduction for high-dimensional data. Principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] are two most popular linear subspace learning methods. In the past, many LDA extensions have been developed to deal with the small sample size problem, but they fail to realize the essential data structures nonlinearly embedded in high-dimensional space. In order to overcome this limitation, some known manifold learning methods are presented such as neighborhood preserving embedding (NPE) [3], locality preserving projection (LPP) [4], local linear embedding (LLE) [5], local Fisher discriminant analysis (LFDA) [6], Laplacianfaces [7] and unsupervised discriminant projection (UDP) [8], marginal Fisher analysis (MFA) [9], linear discriminant projection (LDP) [10], graph-optimized locality preserving projections (GoLPP) [11] graph-based Fisher analysis [12], hypergraph analysis approach [13,14] and optimized multigraph-based semi-supervised learning (OMG-SSL) [15].

In recent years, sparse representation (or sparse coding) has been attracting a lot of attention due to its great success in image processing, and it has been used for face recognition and texture classification. Transform-invariant sparse representation [16] recovers the sparse representation of a target image and the image plane transformation between the target and the model images simultaneously. Wright et al. [17,18] presented a sparse representation-based classification (SRC) method and successfully applied it to recognize human faces with varying lighting condition, occlusion and disguise. Yang et al. [19] applied the sparse representation for face recognition with occlusion based on the Gabor feature. Meanwhile, some known feature extraction algorithms based on sparse learning are proposed, sparse principal component analysis [20] uses Lasso (elastic net) to produce modified principal components with sparse learning; sparse projection [21] based on a graph embedding model learns a set of sparse basis function by applying regularized regression; sparse preserving projection [22,23] aims to preserve the sparse reconstructive relationship of the data by minimizing a regularization-related objection function; Yang and Chu [24] used the decision rule of SRC to steer the design of a dimensionality reduction method, i.e. the sparse representation classifier steered discriminative projection (SRC-DP); Zhang et al. [25] proposed a novel linear subspace learning approach via sparse coding.

Although the above methods have shown success in classification and feature extraction, there are some limitations as follows:

1. The feature extraction and classification are separate. In [1–12], the feature extraction algorithms are firstly used to reduce the dimensionality of high-dimensional data, than classifiers are applied to measure the performance of the feature extraction

* Corresponding author. Tel.: +86 25 8431 7297x403.
E-mail address: cuiyan899@163.com (Y. Cui).

http://dx.doi.org/10.1016/j.neucom.2014.03.034
0925-2312/

criterions, similarly, the classifiers are applied to classify the dimension reduced data [13–19] independently. Therefore these methods are implemented through two stage: feature extraction stage and classification stage.
2. SRC cannot be applied to high-dimensional data directly. For high-dimensional data, the sparse representation coefficients are hardly obtained since the dictionary is not over-complete. Therefore, in order to obtain sparse representation coefficients, the dimensionality of data is pre-reduced by random matrix [17–19] and PCA [24,25].
3. The feature extraction criterions proposed in [20,21] and [24,25] cannot extract features directly. In [20,21], a new feature extraction criterion is proposed under the sparse constraint of the projection vectors. In [24,25], the scatter matrices of the samples are redefined based on the results of SRC. In order to obtain the sparse projection vectors and the representation coefficients, PCA is used to preprocess data, rather than extract features directly, so these criterions are equivalent to two stage feature extraction.
4. The entire training samples used as dictionary may affect the performance of sparse representation. In [22–26], the original image samples are used to represent the input data, actually the original training samples have much redundancy as well as noise and trivial information that can be negative to the recognition. In addition, if the training samples are huge, the computation of the sparse representation will be time consuming, it is needed a more compact and robust dictionary such that each sample in the test set can be represented as a sparse linear combination of its atoms.

In order to overcome the above limitations, in this paper, a novel fusion algorithm, namely feature extraction and classification fusion algorithm (FECFA), is developed to implement feature extraction and classification simultaneously. More specifically, an optimal over-complete dictionary is adaptively learned from the original training data to represent the test data, and the learned optimal dictionary may bate the redundancy as well as noise and trivial information of the original training data. Meanwhile, in order to obtain the sparse representation coefficients, a sparse preserving embedding map is sought to reduce the dimensionality of the test data. Lastly, the test data can be classified by the sparse representation coefficients. For FECFA, need of special note is that the sparse preserving embedding map is learned based on the classification criterions, and the sparse preserving embedding map and the sparse representation coefficients can be obtained by solving an optimization problem alternately. In contrast to the state-of-the-art feature extraction and classification methods, FECFA has the following advantages:

1. FECFA can reduce the dimensionality of a query sample and classify it simultaneously. Therefore, the features of the high-dimensional data need not be pre-processed before classification, and the feature extraction method need not classifier to measure its performance.
2. The sparse preserving embedding map is learned from test data and training data, rather than from training data only. So FECFA utilizes the prior knowledge of test data, which may be practical in the real application. Furthermore, the high-dimensional data pre-processed by the sparse preserving embedding map will improve positive recognition of classification because the sparse preserving embedding map is learned based on the classification criterions.
3. Through sparse preserving, we need not decide how many k-nearest neighbors to be selected to reconstruct the samples such as in [5–11], so FECFA is more adaptive.
4. The test data is represented by the learned optimal dictionary. In contrast to the original sample, the learned dictionary may bate the redundancy as well as noise and trivial information of the original training data.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, a novel supervised fusion algorithm for feature extraction and classification is proposed. Experiments are presented in Section 4. Conclusions are summarized in Section 5.

2. Brief review of the related work

In this section, we introduce the basic idea of the Neighborhood Preserving Embedding (NPE), sparse representation-based classifier and k-SVD dictionary learning algorithm.

2.1. The neighborhood preserving embedding

NPE is an unsupervised manifold learning algorithm that computes low-dimensional, neighbor-hood-preserving embedding of high-dimensional inputs. Specifically, we expect data point and its neighbors to lie on or close to a locally linear patch of the manifold and the local reconstruction errors of these patches are measured by

$$e(w)=\sum_i \| x_i-\sum_{j=1}^{k} w_{ij}x_j \|_2^2. \tag{1}$$

Suppose that the data lies on or near a smooth nonlinear manifold of lower dimensionality $d \ll m$, NPE constructs a neighborhood-preserving mapping P to map the high-dimensional observation x_i to a low dimensional vector y_i, where y_i represents global internal coordinates on the manifold. So in the low dimensional space, Eq. (1) becomes

$$e(w)=\sum_i \| y_i-\sum_{j=1}^{k} w_{ij}y_j \|_2^2=\sum_i \| P^Tx_i-\sum_{j=1}^{k} w_{ij}P^Tx_j \|_2^2 \tag{2}$$

where $y_i=P^Tx_i, (i=1,2,\ldots,n)$. Once neighbors are chosen, the optimal weights w_{ij} and the neighborhood-preserving mapping P are computed by standard linear algorithm.

2.2. The sparse representation-based classifier

The sparse representation-based classifier (SRC), which adaptively chooses the minimal number of training samples to represent each test sample, can be considered as a generalization of nearest neighbor (NN) [28] and nearest subspace (NS) [29]. Suppose that there are c known pattern classes and $A_i=[x_{i1},x_{i2},\ldots,x_{in_i}]\in R^{m\times n_i}, (i=1,2,\ldots,c)$ is the data matrix formed by i-th class samples, any new (test) sample $z\in R^m$ from i-th class will be approximately represented as a sparse linear combination of i-th class samples as follows:

$$z=\alpha_{i1}x_{i1}+\alpha_{i2}x_{i2}+\cdots+\alpha_{in_i}x_{in_i} \tag{3}$$

where $\alpha_{ij}\in R, (j=1,2,\ldots,n_i)$. Let us define a new matrix A for the entire training samples

$$\begin{aligned} A&=[A_1,A_2,\ldots,A_c]\\ &=[x_{11},x_{12},\ldots,x_{1n_1},\ldots,x_{i1},x_{i2},\ldots,x_{in_i},\ldots,x_{c1},x_{c2},\ldots,x_{cn_c}]\in R^{m\times n} \end{aligned} \tag{4}$$

where $\sum_{i=1}^{c} n_i=n$. Any new (test) sample y can be represented as a sparse linear combination of all training samples as

$$y=A\alpha\in R^m \tag{5}$$

where $\alpha=[\alpha_{11},\alpha_{12},\ldots,\alpha_{1n_1},\ldots,\alpha_{i1},\alpha_{i2},\ldots,\alpha_{in_i},\ldots,\alpha_{c1},\alpha_{c2},\ldots,\alpha_{cn_c}]^T\in R^n$ is a coefficient vector. This system of linear equation is undetermined if $m<n$, so its solution is not unique, the sparsest solution can be sought by solving the following optimization problem:

$$(L_0)\quad \alpha^*=\arg\min\|\alpha\|_0,\quad \text{s.t.}\quad A\alpha=y \tag{6}$$

where $\|\alpha\|_0$ denotes the L_0-norm which counts the number of nonzero entries in a vector. Recent research efforts reveal that if the solution α^* is sparse enough, the solution of the L_0-minimization problem of Eq. (6) is approximated to the solution of the following L_1-minimization problem [30–32]:

$$(L_1) \quad \alpha^* = \arg\min \|\alpha\|_1, \quad \text{s.t.} \quad A\alpha = y. \tag{7}$$

This problem can be solved by standard linear programming method [31]. For each class i, let $\delta_i(\alpha^*)$ be a vector whose only nonzero entries are the entries in α^* that are associated with i-th class. Using only the coefficients associated with i-th class, one can reconstruct the given test sample y as $\hat{y}_i = A\delta_i(\alpha^*)$. The SRC decision rule is assigning it to the object class that minimizes the residual between y and $\hat{y}_i$, i.e.

$$\min_i r_i(y) = \|y - A\delta_i(\alpha^*)\|_2^2. \tag{8}$$

2.3. Dictionary selection

In this section, we review the k-SVD algorithm designed overcomplete dictionaries for sparse representation [27]. The k-SVD algorithm finds the best dictionary $D = [d_1, d_2, \ldots, d_k]$ to represent the data samples $\{x_i\}_{i=1}^n$ as sparse composition by two stages.

Firstly, sparse coding stage: use any pursuit algorithm to compute the representation coefficient vector α^* for each sample x_i by approximating the solution of the following optimization problem:

$$\min_{\alpha_i} \|x_i - D\alpha^i\|_2^2 \quad \text{s.t. } \|\alpha^i\|_0 \le T_0 \quad (i = 1, 2, \ldots, n). \tag{9}$$

Secondly, codebook update stage: for each column $k = 1, 2, \ldots, k$ in $D^{(J-1)}$, update it by the following steps:

Step 1: Define the group of samples that use this atom $w_k = \{i | 1 \le i \le n, \alpha_T^k(i) \ne 0\}$, where α_T^k is the coefficients that corresponding to the atom d_k in the dictionary.

Step 2: Compute the overall representation error matrix E_k by $E_k = Y - \sum_{j \ne k} d_j x$.

Step 3: Restrict E_k by choosing only the columns corresponding to those elements that initially used d_k in their representation, and obtain E_k^R.

Step 4: Apply SVD decomposition $E_k^R = U\Delta V^T$, and choose the updated dictionary column $\tilde{d}_i$ to be the first column of U, update the coefficient vector α_k^R to be the first column of V multiplied by $\Delta(1,1)$. For details, see [27].

3. The novel supervised feature extraction and classification fusion algorithm

In this section, we introduce the basic ideas of FECFA for feature extraction and classification. Specifically, an optimal dictionary is learned from the training data set to represent the test data; then a sparse preserving embedding map and sparse coefficients are optimized alternately. From the sparse preserving embedding map and sparse coefficients, we can reduce the dimensionality of the test data and classify it simultaneously. Furthermore, we generalize FECFA to the incremental test data set.

3.1. The novel supervised feature extraction and classification fusion algorithm for fixed test data

In this section, we will incorporate the class information to construct a fusion algorithm for feature extraction and classification. Suppose there are c known pattern classes and $X_i = [x_{i1}, x_{i2}, \ldots, x_{in_i}] \in R^{m \times n_i}$, $(i = 1, 2, \ldots, c)$ is the i-th class training samples matrix. Let us define a matrix $X = [X_1, X_2, \ldots, X_c] \in R^{m \times n}$, where $n = \sum_{i=1}^c n_i$. The matrix X is obviously composed of entire training samples. A test data $y \in R^m$ can be well approximated by the linear combination of the training data, i.e.

$$y = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \alpha_{ij} x_{ij} + \varepsilon = X\alpha + \varepsilon \tag{10}$$

where ε is the data noise, $\alpha = [\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1n_1}, \ldots, \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in_i}, \ldots, \alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cn_c}]^T \in R^n$ is the linear combination coefficients vector. Intuitively, if y from i-th class, y will be approximately represented by i-th class samples. Let $\delta_i(\alpha)$ be the representation coefficients vector with respect to i-th class, then the prototype of i-th class with respect to y is $\hat{y}_i = X\delta_i(\alpha)$ and the residual between y and $\hat{y}_i$ is minimized. Meanwhile, the prototype of j-th class $(j \ne i)$ is $\overline{y}_j = X\delta_j(\alpha), (j \ne i)$, where $\delta_j(\alpha)$ is the representation coefficients vector with respect to j-th class, and the residual between y and $\overline{y}_j$ is larger than that between y and $\hat{y}_i$. To make SRC achieve good performance, we expect the within class residual minimized, while the sum of the between class residual maximized, simultaneously. Therefore the optimization problem is formed

$$\min_\alpha \|y - X\delta_i(\alpha)\|_2^2 - \sum_{j \ne i} \|y - X\delta_j(\alpha)\|_2^2 + \lambda\|\alpha\|_1. \tag{11}$$

Let $\delta_{\bar{i}}(\alpha)$ is the representation coefficients vector with respect to non-i-th class, i.e. its only nonzero entries are the entries in α that are associated with j-th's class $(j \ne i)$, Eq. (11) can be converted into

$$\min_\alpha \|y - X\delta_i(\alpha)\|_2^2 - \|y - X\delta_{\bar{i}}(\alpha)\|_2^2 + \lambda\|\alpha\|_1. \tag{12}$$

However, the original training samples set X has much redundancy as well as noise and trivial information that can be negative to the recognition. In addition, if the training samples are huge, the algorithm will be time consuming, so an optimal dictionary is needed for sparse representation. Therefore, we use k-SVD to adaptively learn a dictionary $D \subset X$ from the following cases:

1. Shared dictionary – learning a single dictionary $D = \{d_1, d_2, \ldots, d_k\} \subset X$ from the training data, when the training data is not abundant.
2. Concatenated dictionaries – learning $D_i = \{d_{i1}, d_{i2}, \ldots, d_{ik_i}\} \subset X_i$, $(i = 1, 2, \ldots, c)$ for i-th class samples, and concatenating $D_i, (i = 1, 2, \ldots, c)$ as a single dictionary $D = \{D_1, D_2, \ldots, D_c\} \subset X$, when the training data is abundant.

Based on the learned optimal dictionary, Eq. (12) becomes

$$\min_\alpha \|y - D\delta_i(\vartheta)\|_2^2 - \|y - D\delta_{\bar{i}}(\vartheta)\|_2^2 + \lambda\|\vartheta\|_1 \tag{13}$$

where $\vartheta = [\vartheta_1, \vartheta_2, \ldots, \vartheta_k]^T$ is the linear combination coefficients vector with the learned dictionary D, $\delta_k(\vartheta)(k = i, j)$ is the representation coefficients vector with respect to k-th class. Furthermore, in order to obtain the sparse representation coefficients vector ϑ, a sparse preserving embedding map $W = [w_1, w_2, \ldots, w_d] \in R^{m \times d}$ needs to be learned to reduce the dimensionality of y, resulting in the following optimization problem:

$$\begin{aligned} \min_{W,\vartheta} \quad & \|W^T y - W^T D\delta_i(\vartheta)\|_2^2 - \|W^T y - W^T D\delta_{\bar{i}}(\vartheta)\|_2^2 + \lambda\|\vartheta\|_1 \\ \text{s.t.} \quad & w_k^T w_k = 1, \quad k = 1, 2, \ldots, d \end{aligned} \tag{14}$$

For a given test set $U = \{y_1, y_2, \ldots, y_l\}$, the sparse preserving embedding map and the sparse reconstruction coefficients can be obtained by solving the following optimization problem:

$$\begin{aligned} \min_{W,\Lambda} \quad & \|W^T U - W^T D\hat{\Lambda}\|_F^2 - \|W^T U - W^T D\overline{\Lambda}\|_F^2 + \lambda\|\Lambda\|_1 \\ \text{s.t.} \quad & W^T W = I \end{aligned} \tag{15}$$

where $\hat{\Lambda} = [\hat{\delta}_i(\vartheta^1), \hat{\delta}_i(\vartheta^2), \ldots, \hat{\delta}_i(\vartheta^l)]$ is the linear combination coefficients matrix corresponding the minimum reconstruct error class

samples, $\overline{\Lambda}=[\overline{\delta}_{\overline{i}}(\vartheta^1),\overline{\delta}_{\overline{i}}(\vartheta^2)\cdots,\overline{\delta}_{\overline{i}}(\vartheta^l)]$ is the linear combination coefficients matrix corresponding the non-minimum reconstruct error class, $I\in R^{d\times d}$ is an identity matrix. Eq. (15) is not convex in W and Λ simultaneously, but it is convex in one once the other fixed. Therefore, Eq. (15) can be done in an alternative coordinate descent fashion between W and Λ, which guarantees to converge to a local minimum. Next, we will discuss the algorithm derivation. Firstly, when Λ is fixed, Eq. (15) is equivalent to

$$\min_{W} \quad \|W^TU-W^TD\hat{\Lambda}\|_F^2-\|W^TU-W^TD\overline{\Lambda}\|_F^2$$
$$\text{s.t.} \quad W^TW=I. \tag{16}$$

The above problem can be further written as

$$\min_{W} \quad \mathrm{tr}[W^T(U-D\hat{\Lambda})(U-D\hat{\Lambda})^TW]-\mathrm{tr}[W^T(U-D\overline{\Lambda})(U-D\overline{\Lambda})^TW]$$
$$\text{s.t.} \quad W^TW=I. \tag{17}$$

Let $S_w=(U-D\hat{\Lambda})(U-D\hat{\Lambda})^T$ and $S_b=(U-D\overline{\Lambda})(U-D\overline{\Lambda})^T$, the problem can be converted into

$$\min_{W} \quad \mathrm{tr}(W^T(S_w-S_b)W)$$
$$\text{s.t.} \quad W^TW=I. \tag{18}$$

The above formula is equivalent to

$$\min_{w_k} \quad \sum_{k=1}^{d} w_k^T(S_w-S_b)w_k$$
$$\text{s.t.} \quad w_k^Tw_k=1, \quad k=1,2,\ldots,d. \tag{19}$$

To solve the above optimization problem, a Lagrangian function may be introduced

$$L(w_k,\mu_k)=\sum_{k=1}^{d}[w_k^T(S_w-S_b)w_k]-\mu_k(w_k^Tw_k-1) \tag{20}$$

with multipliers μ_k. The Lagrangian $L(w_k,\mu_k)$ has to be minimized with respect to w_k and μ_k. Taking the derivatives of $L(w_k,\mu_k)$ with respect to w_k, we obtain

$$\frac{\partial(L(w_k,\mu_k))}{\partial(w_k)}=[(S_w-S_b)-\mu_kI]w_k. \tag{21}$$

Let the derivative be zero, and we have

$$(S_w-S_b)w_k=\mu_kw_k, \quad k=1,2,\ldots,d, \tag{22}$$

which means that μ_k's are the eigenvalues of S_w-S_b and w_k's are the corresponding eigenvectors. Therefore, the sparse preserving embedding map W is composed by the d's eigenvectors corresponding the d's least eigenvalues.

Secondly, when W is fixed, Eq. (15) with respect to Λ can be decomposed into l independent l_1-norm regularized regression problem:

$$\min_{\Lambda(:,t)} \quad \|W^TU(:,t)-W^TD\hat{\Lambda}(:,t)\|_F^2$$
$$-\|W^TU(:,t)-W^TD\overline{\Lambda}(:,t)\|_F^2+\lambda\|\Lambda(:,t)\|_1. \tag{23}$$

The above model can be efficiently solved by the Lasso algorithm [32], and $U(:,t)$ can be classified with $\Lambda(:,t)$ based on the SRC decision rule, the main procedures are summarized in Algorithm 1.

From the sparse preserving embedding map and the corresponding coefficients, we can reduce the dimensionality of the high-dimensional data and classify it simultaneously. In contrast to the state-of-the-art sparse represent methods, we use k-SVD to adaptively learn an optimal dictionary, which bates the redundancy as well as noise and trivial information of the original training data. Furthermore, in order to obtain the sparse coefficients, a sparse preserving embedding map is integrated into the classification fusion algorithm instead of pre-reducing the dimensionality of data by PCA independently. In construct to the classic dimensionality reduced methods, the sparse preserving embedding map is learned from the test and training data set, rather than from the training data set only. Therefore the proposed method takes into account the prior knowledge of the test data.

Algorithm 1. Dict-based FECFA.

Step 1: Learn dictionary $D_i(i=1,2,\ldots,c)$ from the i-th class training data by using k-SVD;

Step 2: Obtain the minimal norm least square solution of $\|y_i-D\vartheta^i\|_2^2<\epsilon$ as the initial $\vartheta^i(i=1,2,\ldots,l)$, where $D=[D_1,D_2,\ldots D_c]$ is concatenate matrix;

Step 3: From the initial ϑ^i and $D_i(i=1,2,\ldots,c)$, compute the reconstruct error and seek $\hat{\delta}_i(\vartheta^i)$ and $\overline{\delta}_{\overline{i}}(\vartheta^i)(i=1,2,\ldots,l)$;

Step 4: From the selected dictionary D and the initial $\hat{\Lambda}_0$ and $\overline{\Lambda}_0$, seek the sparse preserving embedding projection by solving the optimization problem Eq. (18), and reduce the dimensionality of the test and training data by the sparse preserving embedding projection;

Step 5: Fixed $J_k=\|W^TU-W^TD\hat{\Lambda}\|_F^2-\|W^TU-W^TD\overline{\Lambda}\|_F^2$, implement alternate iteration method to the optimization problem Eq. (15), until $|J_k-J_{k-1}|<\epsilon$, and seek the optimal solution Λ and classify $U(:,t)$ with $\Lambda(:,t)$ based on the SRC decision rule.

3.2. Incremental learning for the test data

In Section 3.1, we deduce FECFA when the test data is fixed. In the real application, the test data usually becomes available gradually. This fact requires FECFA to have the capability to learn the test data incrementally. For a fix test set $U=\{y_1,y_2,\ldots,y_l\}$, the sparse preserving embedding map $W\in R^{m\times d}$ and the sparse reconstruction coefficient matrix Λ can be sought by Eq. (15). Similarly, for the new incremental test data $U_{new}=\{y_{l+1},y_{l+2},\ldots,y_{l+m}\}$, the new sparse preserving embedding map $W_{new}\in R^{m\times l}$ and the sparse reconstruction coefficients matrix Λ_{new} can be computed by the following optimization problem:

$$\min_{W_{new},\Lambda_{new}} \quad \|W_{new}^TU_{new}-W_{new}^TD\hat{\Lambda}_{new}\|_F^2$$
$$-\|W_{new}^TU_{new}-W_{new}^TD\overline{\Lambda}_{new}\|_F^2+\lambda\|\Lambda_{new}\|_1$$
$$\text{s.t.} \quad W_{new}^TW_{new}=I \tag{24}$$

where D is the same as in Eq. (16), which is learned from the training data set. Similarly, the sparse preserving embedding map W_{new} and the sparse reconstruction coefficients matrix Λ_{new} can be sought by Algorithm 1.

4. Experiments

In this section, we systematically apply FECFA on two land cover databases. The first land cover database created in 2012 by the Nanjing University of Science and Technology (NJUST), which is composed of 6 classes of land cover (such as dirt roads, sandy roads, tree, vegetation (green), water and vegetation (yellow)), contains 12,000 cropped images with 16×16 pixels derived from six different road condition video files. The second land cover database derived from the Outex Texture Database, which is composed of 5 classes of land cover (such as tree, bushes, grass, roads and buildings) and sky, contains 10,000 cropped images with 64×64 pixels derived from 48 natural scene pictures. Some images of the two databases can be found in Fig. 1. In our experiments, we illustrate the effect of dictionary selection and compare FECFA with classification and feature extraction methods. All the results are performed on Pentium 2.52 GH with 2 G RAM

Fig. 1. Some images of the two databases. (a) Some images of NJUST databases. (b) Some images of the Outex Texture Database.

and programmed in the MATLAB language (version R2011b), the mean positive recognition rate and standard deviation stand are used for performance measure after 5-fold cross validation.

4.1. The effect of dictionary selection

In this section, we illustrate the effect of dictionary selection from two cases. When every class training data is abundant, a dictionary is learned from every class samples to concatenate as a needed dictionary; when the training data is not abundant, a shared dictionary is learned from the training data. For NJUST databases, the different size of dictionary is selected from every class samples to test the effect of dictionary in the *k*-SVD dictionary learning. Specifically, every class samples are randomly split to the train set and test set with the ratio 4:1, and the different size of sub-dictionary is selected from every class samples to concatenate the whole dictionary to represent the test samples. Then, FECFA is applied to reduce the dimensionality and classification based on the learned dictionary simultaneously. The main results of the first land cover databases can be found in Table 1.

However, the different size of dictionary cannot be obtained from every class samples when the corresponding class samples is not abundant. Thus the dictionary should be learned from all class samples. Next, we will use the Outex Texture Database to discuss the effect of the dictionary size from all class samples, samples are randomly split to the train set and test set with the ratio 4:1. The different sizes of dictionary are selected from the training samples to represent the testing samples. Then, FECFA is applied to reduce the dimensionality and classification based on the learned dictionary simultaneously. The main results of Outex Texture Database can be found in Table 2.

According to Tables 1 and 2, we know that the size of the learned dictionary affects the *k*-SVD dictionary learning whether learning the dictionary from every class or learning the dictionary from all class samples. when all samples are selected as dictionary, the mean positive recognition rate does not reach the maximum, so the learned dictionary may bate the redundancy as well as noise and trivial information of the original training data to some extent.

4.2. Comparison with classifiers

In this section, we compare FECFA with random forest (RF), support vector machine (SVM), and SRC on three subsets of the Outex Texture Databases. Specifically, every class samples are randomly split to the train set and test set with the ratio 4:1, and then we apply FECFA to reduce the dimensionality and classification based on the learned optimal dictionary. In our experiments, we set 50, 60 and 70 nodes in RF and the linear kernel and polynomial kernel in SVM. The main results can be found in Table 3.

Table 1
The mean positive recognition rate (%) and standard deviation with respect to color feature (case 1: dirt roads, tree, vegetation (green), vegetation (yellow), sandy roads, water; case 2: water, dirt roads, sandy roads; case 3: tree, vegetation (green), vegetation (yellow)).

Dict. size ratio	768/1600	1024/1600	1280/1600	1536/1600
Case 1	68.18 ± 0.65	68.73 ± 0.41	$\mathbf{69.45 \pm 0.60}$	69.43 ± 0.34
Case 2	66.23 ± 0.59	66.50 ± 0.36	$\mathbf{66.77 \pm 0.96}$	66.73 ± 0.54
Case 3	81.83 ± 0.94	$\mathbf{82.82 \pm 0.32}$	82.62 ± 1.01	82.47 ± 0.99

According to Table 3, we have the following conclusion: (1) for all subset, the dimensionality of the original data can be obviously reduced by FECFA, thus storage space can be saved to a great extent. (2) For the mean positive recognition rate, FECFA is slightly larger than others classifier with respect to color feature except for 5 class; for the LBP texture feature, FECFA is significantly superior to SVM and slightly better than SRC for 4 class and 3 class data subset, but weaker than RF for 4 class and 3 class data subset. (3) For the standard deviation, FECFA is significantly superior to RF, SVM and SRC with respect to color feature and as similar as RF, SVM and SRC for the LBP texture feature. In a word, FECFA is similar to SRC and RF and is obviously superior to SVM.

4.3. Comparison with feature extraction

In this section, FECFA is compared with feature extraction algorithms. For LPP and LDA, the feature extraction algorithms are firstly applied to reduce the dimensionality of the test data, and then SRC is used to classify it. While FECFA is directly applied to reduce the dimensionality of the test data and classify it simultaneously. The main results can be found in Fig. 2.

According to Fig. 2, we can find that the mean positive recognition rate of FECFA is significantly larger than that of LPP+SRC and LDA+SRC for the two databases under different dimensionality except for the dimensionality of the Outex texture databases is 31.

5. Conclusion

In this paper, we present a novel supervised feature extraction and classification fusion algorithm based on dictionary selection. According to Section 4.1, we know that the size of dictionary affects the positive recognition rate of the proposed fusion algorithm and the learned dictionary may bate the redundancy as well as noise and trivial information. According to Section 4.2, for the all cropped images databases, the dimensionality of the original data can be reduced by the proposed fusion algorithm obviously; and the proposed fusion algorithm is superior to RF, SVM and SRC.

Table 2
The mean positive recognition rate (%) and standard deviation with respect to LBP texture feature (case 1: sky, tree, grass, roads, bushes, buildings; case 2: tree, grass, roads, bushes; # : no data).

Dict size ratio	453/1422	604/1422	755/1422	906/1422	1057/1422	1208/1422	1422/1422
Case 1	74.41 ± 1.08	78.12 ± 0.96	80.92 ± 1.25	81.54 ± 1.35	85.86 ± 1.23	**87.94 ± 0.48**	85.02 ± 0.95
Dict size ratio	453/948	604/948	755/948	906/948	948/948	#	#
Case 2	80.89 ± 2.93	82.07 ± 1.65	83.09 ± 1.64	**83.83 ± 1.07**	82.58 ± 2.56	#	#

Table 3
The mean positive recognition rate (%) and standard deviation of the Outex Texture databases. (·) represents dimension (from the first line to third line, the node is 50, 60 and 70 w.r.t. RF; from the fifth to sixth line, the linear kernel and polynomial kernel w.r.t. SVM).

Method	Color feature			LBP texture feature		
	Sky, tree, grass, roads, buildings	Sky, tree, grass, roads	Sky, tree, grass	Sky, tree, grass, roads,buildings	Sky, tree, grass, roads	Sky, tree, grass
RF	96.39 ± 0.75 (512)	99.44 ± 0.28 (512)	99.34 ± 0.22 (512)	91.95 ± 0.74 (512)	94.82 ± 1.02 (512)	96.10 ± 1.03 (512)
	97.58 ± 0.22 (512)	99.09 ± 0.31 (512)	98.86 ± 0.43(512)	91.61 ± 0.89 (512)	93.83 ± 0.58 (512)	95.77 ± 0.67 (512)
	97.65 ± 0.17 (512)	99.06 ± 0.26 (512)	98.86 ± 0.43(512)	91.82 ± 0.76 (512)	94.11 ± 0.84 (512)	95.66 ± 1.04 (512)
SVM	96.32 ± 0.88 (512)	98.65 ± 0.56 (512)	98.38 ± 0.75 (512)	77.47 ± 2.60 (512)	81.88 ± 2.38 (512)	82.72 ± 2.37 (512)
	86.45 ± 1.21 (512)	92.11 ± 1.15 (512)	92.00 ± 0.70 (512)	67.59 ± 1.81 (512)	74.71 ± 2.35 (512)	80.11 ± 1.77 (512)
SRC	98.97 ± 0.33 (512)	99.54 ± 0.21 (512)	99.39 ± 0.29 (512)	92.84 ± 0.79 (512)	93.09 ± 0.68 (512)	93.60 ± 0.87 (512)
FECFA	98.87 ± 0.32 (72)	99.64 ± 0.14 (29)	99.52 ± 0.18 (29)	92.84 ± 0.92 (42)	94.08 ± 0.42 (43)	94.82 ± 1.02 (20)

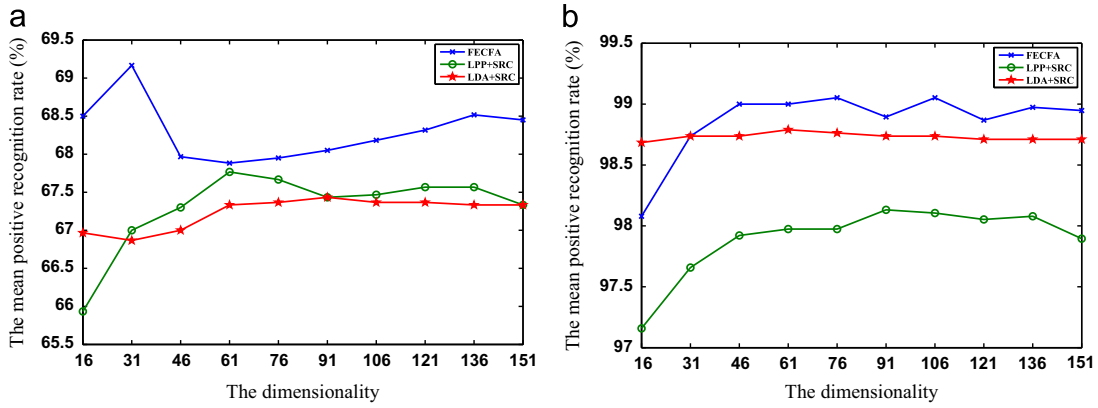


Fig. 2. The mean positive recognition rate with respect to color feature. (a) The NJUST databases, (b) The Outex texture databases.

According to Section 4.3, we know that the proposed fusion algorithm is superior to LPP and LDA under different dimensionality. Therefore the presented fusion algorithm can fuse feature extraction and classification excellently. In order to test its generalization and stability, we will apply the fusion algorithm to microarray gene expression data, information retrieval, web document classification and etc, and kernelize the fusion algorithm for nonlinear data set.

Acknowledgments

This work is supported by the National Science Foundation of China (Grant no. 61233011, 61373063, 613750071, 91220301, 61125305 and 61005005), the National Science Fund for Distinguished Young Scholars (Grant no. 61125305) and the Doctoral Candidate Creative Fund of Jiangsu province (Grant no. CXZZ12_0207).

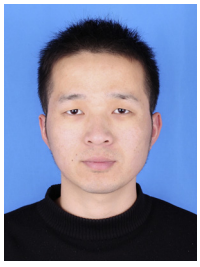
References

[1] I. T Jolliffe, Principal Component Analysis, second ed., Springer-Verlag, New York, 2002.
[2] S. Mika, Kernel Fisher discriminant (Ph.D. thesis), University of Technology, Berlin, 2002.
[3] X.F. He, D. Cai, Sh.Ch. Yan, H.J. Zhang, Neighborhood preserving embedding, in: Proceedings of the 2005 IEEE International Conference on Computer Vision, 2005, pp. 1208–1213.
[4] X.F. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems, vol. 16, 2003.
[5] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[6] M. Sugiyama, Local Fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 905–912.
[7] X.F. He, Sh.Ch. Yan, Y.X. Hu, P. Niyogi, H. J Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[8] J. Yang, D. Zhang, J.Y. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 650–664.
[9] Sh.Ch. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[10] H.P. Cai, K. Mikolajczyk, J. Matas, Learning linear discriminant projections for dimensionality reduction of image descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 338–352.
[11] L.M. Zhang, L.Sh. Qiao, S.C. Chen, Graph-optimized locality preserving projections, Pattern Recognit. 43 (6) (2010) 1993–2002.
[12] Y. Cui, L.Y. Fan, A novel supervised dimensionality reduction algorithm: graph-based Fisher analysis, Pattern Recognit. 45 (4) (2012) 1471–1481.
[13] Y. Gao, M. Wang, D.Ch. Tao, R.R. Ji, Q.H. Dai, 3-D object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process., Hypergraph Anal. 21 (9) (2012) 4290–4303.
[14] Y. Gao, M. Wang, Zh.J. Zha, J.L. Shen, X.L. Li, X.D. Wu, Visual-textual joint relevance learning for tag-based social image search, IEEE Trans. Image Process. 22 (1) (2013) 363–376.

[15] M. Wang, X.Sh. Hua, R.ch. Hong, J.H. Tang, G.J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009) 733–746.
[16] J.Zh. Huang, X.L. Huang, D. Metaxas, Simultaneous image transformation and sparse representation recovery, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[17] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[18] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, Sh.Ch. Yan, Sparse representation for computer vision and pattern recognition, in: Proceedings of the IEEE, Special Issue on Applications of Compressive Sensing and Sparse Representation, vol. 98 (6), 2010, pp. 1031–1044.
[19] M. Yang, L. Zhang, Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary, Lect. Notes Comput. Sci. 6316 (2010) 448–461.
[20] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.
[21] D. Cai, X.F He, J.W Han, Sparse projection over graph, in: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008, pp. 610–615.
[22] L.Sh. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving projection with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.
[23] L.Sh. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving discriminant analysis for single training image face recognition, Pattern Recognit. 31 (5) (2010) 422–429.
[24] J. Yang, D. Chu, Sparse representation classifier steered discriminative projection, in: Proceedings of the Twentieth International Conference on Pattern Recognition, 2010, pp. 694–697.
[25] L. Zhang, P.F. Zhu, Q.H. Hu, D. Zhang, A linear subspace learning approach via sparse coding, in: Proceedings of the 2011 IEEE International Conference on Computer Vision, 2011, pp. 755–761.
[26] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Signal Process. 17 (1) (2008) 53–69.
[27] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[28] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.
[29] S.Z. Li, J.W. Lu, Face recognition using the nearest feature line method, IEEE Trans. Neural Netw. 10 (2) (1999) 439–443.
[30] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Supervised dictionary learning, in: Advances in Neural Information Processing Systems, vol. 21, 2009.
[31] K.K. Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, T.J. Sejnowskl, Dictionary learning algorithms for sparse representation, Neural Comput. 15 (2) (2006) 349–396.
[32] D.L. Donoho, For most large underdetermined systems of linear equations the minimal L_1-norm solution is also the sparsest solution, Commun. Pure Appl. Math. 59 (6) (2006) 797–829.

Yan Cui was born in Shandong, China, in 1985. She received her B.S. and M.S. degrees from Liaocheng University, Liaocheng, China, in 2008 and 2011, respectively. Now she is studying for her Ph.D. degree in pattern recognition and intelligence system from School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. She visited the Department of Electrical and Computer Engineering, University of Miami, USA, from May 2013 to November 2013; she is currently working as a research assistant at the Institute of Textiles and Clothing, Hong Kong Polytechnic University. Her current research interests include pattern recognition, machine learning and image retrieval.

Zhong Jin received the B.S. degree in Mathematics, the M.S. degree in Applied Mathematics and the Ph.D. degree in pattern recognition and intelligence system from Nanjing university of Science and Technology (NUST), Nanjing, China, in 1982, 1984, and 1999, respectively. He is a professor in the School of Computer Science and Engineering, NUST. He had a stay of 15 months as a research assistant at the Department of Computer Science and engineering, the Chinese University of Hong Kong from 2000 to 2001. He visited the Laboratoire HEUDIASYC, Universite de Technologie de Compiegne, France, from October 2001 to July 2002. He visited the Centre de Visio per Computador, Universitat Autonoma de Barcelona, Spain, as the Ramon y Cajal program Research Fellow from September 2005 to October 2005. His current interests are in the areas of pattern recognition, computer vision, face recognition, facial expression analysis and content-based image retrieval.



Jielin Jiang received the B.S. degree from Huainan Normal University, Huainan, China in 2007, the M.S. degree from Inner Mongolia University of Technology, Hohhot, China in 2010. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. From February 2013 to August 2013, he was an Exchange Student with the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. His current research interests include image denoising and image classification.