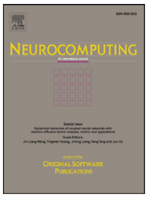
Contents lists available at ScienceDirect

Neurocomputing

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



Hierarchical sparse representation with deep dictionary for multi-modal classification

Zhengxia Wang[a,b], Shenghua Teng[c,*], Guodong Liu[d], Zengshun Zhao[c], Hongli Wu[e,*]

[a] *Department of information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[b] *Department of Automation, Chongqing University, Chongqing 400044, China*
[c] *College of Electronics, Communication and Physics, Shandong University of Science and Technology, Shandong 266590, China*
[d] *College of Civil Engineering, Chongqing Jiaotong University, Chongqing 400074, China*
[e] *College of Information Science and Technology, Hainan Normal University, Hainan 571100, China*

ARTICLE INFO

Article history:
Received 14 May 2016
Revised 2 November 2016
Accepted 16 November 2016
Available online 8 March 2017

Keywords:
Hierarchical sparse representation
Deep dictionary
Multi-view feature

ABSTRACT

Sparse representation based classification (SRC) methods have achieved many successes in pattern recognition and machine learning. In such methods, the training samples of all categories are mixed and compose a dictionary to represent the test sample via sparsity constraint. Then, the class with the minimum representation error wins for labeling the test sample. In general, SRC is more flexible and effective than many supervised learning methods. However, in some cases it is unlikely to represent the test sample accurately, which tends to undermine the classification accuracy. To alleviate this issue, we propose a hierarchical sparse representation based classification method by augmenting the single-layer sparse representation into the hierarchical representation with a deep dictionary. Specifically, the features from all training samples are first divided into several groups according to their labels. Then we employ hierarchical clustering in each group and combine them to form a deep dictionary such that the root layer includes only a certain amount of the most representative exemplars while the subsequent layers focus on characterizing the remaining individual information across different groups. Furthermore, we use the layer-after-layer residuals to encode the variation patterns across individuals in different scales. Given the deep dictionary, a hierarchical sparse representation based classification method is presented to determine the label for each test sample by iteratively representing its primary part with the exemplars in different groups but the remaining parts by the variation patterns encoded in different layers. To further improve the classification accuracy and robustness, we extend our method by taking advantage of the complementary information in multi-view features. Experiments on Multiple Features Data Set show promising results compared with the state-of-the-art classification methods.

1. Introduction

Advanced statistical machine learning and pattern recognition techniques have been actively applied to image analysis and classification [1–4]. Sparse representation based classification (SRC) has achieved a great of interest in classification and feature selection [5–7]. In SRC, features from all training data are mixed to form a feature dictionary. Then, the new observation is sparsely represented by the dictionary, where the atoms of different classes in the dictionary contribute differently. And finally the class label of test sample is determined by the training class which has the smallest residual in representing the test sample. SRC has achieved great success, and some of its extensions have also been developed [8–10]. However, the performance of SRC would deteriorate if the feature dimensionality is too high, or the number of dictionary atoms is too low, or the data is very noisy. Under these circumstances, the representation power of the conventionally simple dictionary will decrease and some test samples cannot be represented accurately.

To address this issue, some methods have been proposed to incorporate intra-class variations as the supplement to form an augmented dictionary [11,12], or adopting dictionary learning for a more discriminative dictionary [13,14]. The use of intra-class variations can enrich the details of the dictionary for sample representation; however, some samples still cannot be well represented with this kind of traditionally “flat” dictionary.

In this respect, we present the hierarchical sparse representation based classification (HSRC) by using a deep dictionary, where the information of entire training dataset is hierarchically encoded with multiple layers. Rather than letting all features extracted

[*] Corresponding authors.
E-mail addresses: tengshenghua@163.com (S. Teng), 18976176119@189.cn (H. Wu).

http://dx.doi.org/10.1016/j.neucom.2016.11.079
0925-2312/© 2017 Elsevier B.V. All rights reserved.

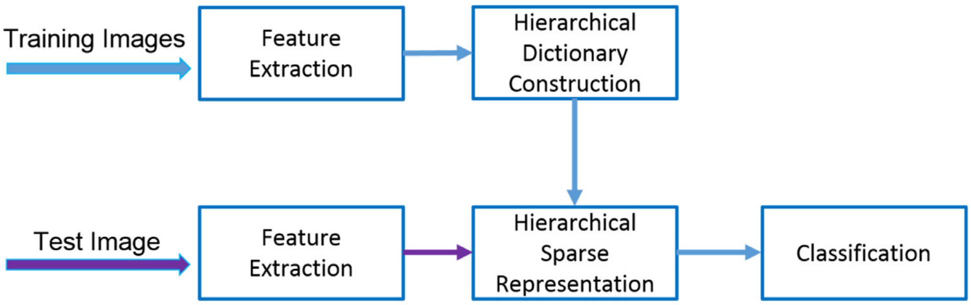


Fig. 1. Flowchart of the hierarchical sparse representation based classification.

from training data mixed in the single-layer dictionary, we propose to construct the deep dictionary to make the sparse representation more efficient. Specifically, we first extract features from each training sample and then divide all features into several groups based on their labels. In particular, affinity propagation (AP) clustering [15] is hierarchically performed to treat sample features differently in a major to minor manner. Therefore, the root layer only comprises a small number of most representative exemplars in each group. The next layer includes the instances with less majority patterns within the same group and so on. Note that, instead of using the original features in the root layer, for each subsequent non-root layer we compute the element-wise residual for each training sample with respect to its parent cluster in the previous layer to encode just the variation patterns. After that, we construct the deep dictionary by merging the clustering results across categories in each layer. In the testing stage, the test sample is also hierarchically represented by the deep dictionary, i.e., iteratively represented by principal patterns in the first layer of dictionary and then by a set of lay-after-layer variation patterns in the non-root layers. Finally, we tag the label to the test sample by the group with the smallest reconstruction error. The flowchart of our proposed is shown in Fig. 1.

2. Method

In the conventional SRC method, usually the training data of all groups is directly stacked to compose a single-layer dictionary. The sparse representation for a test sample is conducted estimating a small number of non-zero coefficients for the selected training samples in the dictionary. But, due to the high dimensionality and also the limited number of training samples in the dictionary, the conventional SRC has limited power in representing the test sample by using the single-layer dictionary. In this paper, we propose a hierarchical sparse representation based classification by using the deep (multi-layer) dictionary, as detailed below.

Assume that the feature length is L for each training sample, and we have $N = N_1 +, \dots, N_i +, \dots, N_C$ training samples belonging to totally C groups. The features from each group compose the feature pools, i.e., $\boldsymbol{X}_1, \dots, \boldsymbol{X}_i, \dots, \boldsymbol{X}_C$, where each group of feature vectors, $\boldsymbol{X}_i = \{\vec{\boldsymbol{x}}_i(s) | s = 1, \dots, N_i\}$, consists of N_i features vectors from N_i training samples in the group G_i. Following the same feature extraction procedure, we can also obtain the feature vector for the test sample, denoted by $\vec{y}$.

2.1. Sparse representation based classification (SRC)

In the conventional SRC methods [8], the feature vectors of all training samples are mixed together to form a single-layer dictionary $\boldsymbol{X} = [\boldsymbol{X}_i]_{i=1,\dots,C}$ by arranging all feature vectors column by column. To classify the test sample, we seek for sparse coefficients $\vec{\boldsymbol{\alpha}} = [\vec{\boldsymbol{\alpha}}_1, \dots, \vec{\boldsymbol{\alpha}}_i, \dots, \vec{\boldsymbol{\alpha}}_C]$for the feature vector $\vec{y}$ of test sample, where the length of column coefficient vector $\boldsymbol{\alpha}$ equals to the total number of atoms in the dictionary $\boldsymbol{X}$. The objective function can be formulated as:

$$\hat{\vec{\boldsymbol{\alpha}}} = \text{are}\min_{\vec{\boldsymbol{\alpha}}} \left\| \vec{\boldsymbol{\alpha}} \right\|_1 s.t. \left\| \vec{y} - \boldsymbol{X}\vec{\boldsymbol{\alpha}} \right\|_2^2 \leq \varepsilon \tag{1}$$

Using the estimated sparse coefficients, the reconstruction error of group $G_i (1 \leq i \leq C)$ is calculated as

$$r_i(\vec{\boldsymbol{y}}) = \left\| \vec{\boldsymbol{y}} - \boldsymbol{X}_i \hat{\vec{\boldsymbol{\alpha}}}_i \right\|_2, \quad = 1, \dots, C. \tag{2}$$

In such case, the class with the minimum reconstruction error indicates the label of the test sample via

$$\text{Label}(\vec{\boldsymbol{y}}) = \arg\min_i r_i(\vec{\boldsymbol{y}}). \tag{3}$$

However, the single-layer dictionary $\boldsymbol{X}$ has limited power in representing the test sample $\vec{\boldsymbol{y}}$, thus this non-optimal sparse representation often yields inaccurate classification result.

2.2. Deep dictionary construction

The goal here is to turn the single-layer dictionary $\boldsymbol{X}$ to the deep dictionary $\boldsymbol{D} = \{\boldsymbol{D}^{(h)} | h = 1, \dots, H\}$ with H layers. To construct the first layer of the dictionary, we apply the AP clustering algorithm [15] to find the cluster centers $\boldsymbol{Z}_i^{(1)}$ in the feature pool $\boldsymbol{X}_i$ of each group G_i. It is worth noting that AP clustering is the non-parametric clustering method, which is fully driven by the feature distribution. For each group G_i, the centers of clusters are used to construct the first-layer sub-dictionary $\boldsymbol{D}_i^{(1)}$ by arranging each cluster center in $\boldsymbol{Z}_i^{(1)}$ column by column. Then, the first-layer dictionary $D^1 = [\boldsymbol{D}_1^{(1)}, \dots, \boldsymbol{D}_i^{(1)}, \dots, \boldsymbol{D}_C^{(1)}]$ is constructed from the sub-dictionaries $\boldsymbol{D}_i^{(1)}$ of all C groups in the first layer.

From the second layer, we construct the dictionary $\boldsymbol{D}^{(h)}$ for each non-root layer $h(2 < h \leq H)$ by repeating the following steps: (1) Since each instance in the remaining samples $(\boldsymbol{X}_i - \sum_{h'=1}^{h-1} \boldsymbol{Z}_i^{(h')})$ of group G_i belongs to one of the clusters of $\boldsymbol{Z}_i^{(h-1)}$ in the previous layer, we compute the element-wise residual for each of the remaining sample with respect to its own parent cluster in the previous layer. (2) We apply AP to all of the remaining samples in each group separately to find the new clusters $\boldsymbol{Z}_i^{(h)}$ in the current layer h. (3) We construct the group-specific dictionary $\boldsymbol{D}_i^{(h)}$ in the current layer by arranging the obtained clusters $\boldsymbol{Z}_i^{(h)}$ column by column. (4) We assemble the dictionary $\boldsymbol{D}^{(h)} = [\boldsymbol{D}_i^{(h)}]_{i=1,\dots,C}$ in the current layer by combining the $\boldsymbol{D}_i^{(h)}$ across all different groups.

2.3. Hierarchical sparse representation based on classification (HSRC)

Given the deep dictionary D with Hlayers, the hierarchical representation for the test sample $\vec{y}$ is cooperatively represented by all layers as below:

$$\vec{y} = \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} + \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} + e, \tag{4}$$

where $\vec{\boldsymbol{\alpha}}^{(h)}$ $(h = 1, 2, \ldots, H)$ is the sparse coefficient vector in each layer and e is the reconstruction error. All coefficient vectors here can be computed with L_1 minimization:

$$\begin{aligned} &\left\{\hat{\vec{\boldsymbol{\alpha}}}^{(1)}, \hat{\vec{\boldsymbol{\alpha}}}^{(2)}, \ldots, \hat{\vec{\boldsymbol{\alpha}}}^{(H)}\right\} \\ &\quad = \arg\min_{\{\vec{\boldsymbol{\alpha}}^{(1)}, \vec{\boldsymbol{\alpha}}^{(2)}, \ldots, \vec{\boldsymbol{\alpha}}^{(H)}\}} \lambda_1 \left\| \vec{\boldsymbol{\alpha}}^{(1)} \right\|_1 + \sum_{h=2}^{H} \lambda_h \left\| \vec{\boldsymbol{\alpha}}^{(h)} \right\|_1 \\ &\qquad s.t. \left\| \vec{\boldsymbol{y}} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right\|_2^2 \le \varepsilon. \end{aligned} \tag{5}$$

To solve the minimization problem in Eq. (5), the Augmented Lagrange Multiplier (ALM) method [16] is applied with the following formula:

$$\begin{aligned} &\lambda_1 \left\| \vec{\boldsymbol{\alpha}}^{(1)} \right\|_1 + \sum_{h=2}^{H} \lambda_h \left\| \vec{\boldsymbol{\alpha}}^{(h)} \right\|_1 + \frac{\xi}{2} \left\| \boldsymbol{y} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right\|_2^2 \\ &\quad + \phi^T \left(\boldsymbol{y} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right), \end{aligned} \tag{6}$$

where ϕ is a vector of Lagrange multipliers and ξ is a penalty parameter. We iteratively optimize $\{\hat{\vec{\boldsymbol{\alpha}}}^{(1)}, \hat{\vec{\boldsymbol{\alpha}}}^{(2)}, \ldots, \hat{\vec{\boldsymbol{\alpha}}}^{(H)}\}$ by the following Algorithm 1.

Using these estimated coefficient vectors $\{\hat{\vec{\boldsymbol{\alpha}}}^{(1)}, \hat{\vec{\boldsymbol{\alpha}}}^{(2)}, \ldots, \hat{\vec{\boldsymbol{\alpha}}}^{(H)}\}$, the reconstruction error of the test data in each group can be computed by

$$r_i\left(\vec{\boldsymbol{y}}\right) = \left\| \vec{\boldsymbol{y}} - \boldsymbol{D}_i^{(1)}\hat{\vec{\boldsymbol{\alpha}}}_i^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\hat{\vec{\boldsymbol{\alpha}}}^{(h)} \right\|_2, \quad i = 1, \ldots, C. \tag{7}$$

Note, we only use the sparse coefficients of the underlying group to represent the primary part in group G_i. In the non-root layers, we still allow the variation patterns of all groups to collaboratively represent the remaining part. Finally, the group with the minimal reconstruction error will win for tagging the test sample $\vec{y}$.

Algorithm 1
Algorithm to solve problem (6) by ALM.

```
Input: Deep dictionary D with $H$ layers, test sample $\vec{\boldsymbol{y}}$, parameters $\lambda_h (h{=}1, \ldots, H)$.
output: sparse coefficient $\vec{\boldsymbol{\alpha}}^{(h)} (h = 1, 2, \ldots, H)$
Initialization: $\vec{\boldsymbol{\alpha}}^{(h)}{=}0\ (h{=}1, 2, \ldots, H)$, $\boldsymbol{\phi}{=}0$, $\xi{=}1$, $\xi_{max}{=}10^4$, $\rho{=}1.5$ and $\varepsilon{=}10^{-4}$
While not converge do
    1. Fix others and update $\vec{\boldsymbol{\alpha}}^{(1)}$, by
           $\vec{\boldsymbol{\alpha}}^{(1)} = \arg\min_{\vec{\boldsymbol{\alpha}}^{(1)}} \left\| \vec{\boldsymbol{y}} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right\|_2^2 + \lambda_1 \left\| \vec{\boldsymbol{\alpha}}^{(1)} \right\|_1$
    2. $h = 2$
    3. while $h \le H$ do
         Fix others and update $\vec{\boldsymbol{\alpha}}^{(h)}$ by
         $\vec{\boldsymbol{\alpha}}^{(h)} = \arg\min_{\vec{\boldsymbol{\alpha}}^{(h)}} \left\| \left(\vec{y} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} + \frac{1}{\xi}\boldsymbol{\phi}\right) - \sum_{h'=2, h' \neq h}^{H} \boldsymbol{D}^{(h')}\vec{\boldsymbol{\alpha}}^{(h')} \right\|_2^2 + \sum_{h=2}^{H} \lambda_h \left\| \vec{\boldsymbol{\alpha}}^{(h)} \right\|_1$
         $h = h + 1$;
         end;
    4. Update $\xi$ by $\xi = \min(\xi_{max}, \rho\xi)$.
    5. Update the multipliers $\boldsymbol{\phi} = \boldsymbol{\phi} + \xi \left( \boldsymbol{y} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right)$
    6. Check the converge condition: $\left\| \vec{\boldsymbol{y}} - \boldsymbol{D}^{(1)}\vec{\boldsymbol{\alpha}}^{(1)} - \sum_{h=2}^{H} \boldsymbol{D}^{(h)}\vec{\boldsymbol{\alpha}}^{(h)} \right\|_2^2 \le \varepsilon$.
End
```

Table 1
Multi-view features of handwritten digits.

View	The composition of current view
mfeat-fou view	Contains 76 Fourier coefficients of the character shapes
mfeat-fac view	Contains 216 profile correlations
mfeat-kar view	Contains 64 Karhunen-Love coefficients
mfeat-pix view	Contains 240 pixel averages in 2 × 3 windows
mfeat-zer view	Contains 47 Zernike moments
mfeat-mor view	Contains 6 morphological features

2.4. HSRC on multiple modalities

Information about a pattern in classification can be revealed via different data capture techniques or through different views (such as profile correlations and Fourier coefficients of the character shapes of handwritten digits). Any single method almost cannot provide complete understanding of a certain pattern, so, multi-modal (multi-view) data is becoming more and more popular. From this point of view, extending the HSRC method to multi-modal data will be expected to effectively utilize the complementary information and further improve the classification accuracy and robustness [17].

Suppose we have M modalities, an H-layer dictionary is constructed for each one, denoted as $\boldsymbol{D}_m = \{\boldsymbol{D}_m^{(h)} | h = 1, \ldots, H, m = 1, \ldots, M\}$, respectively. For the m-th modality, we can iteratively optimize the sparse coefficients $\{\hat{\vec{\boldsymbol{\alpha}}}_m^{(1)}, \hat{\vec{\boldsymbol{\alpha}}}_m^{(2)}, \ldots, \hat{\vec{\boldsymbol{\alpha}}}_m^{(H)}\}$ for the test sample. Then the reconstruction errors corresponding to different groups can be computed with the formula

$$\begin{aligned} r_i(\vec{\boldsymbol{y}}) = \sum_{m=1}^{M} \omega_m \left\| \vec{\boldsymbol{y}}_m - \boldsymbol{D}_{im}^{(1)}\hat{\vec{\boldsymbol{\alpha}}}_{im}^{(1)} - \sum\nolimits_{h=1}^{H} \boldsymbol{D}_m^{(h)}\hat{\vec{\boldsymbol{\alpha}}}_{im}^{(h)} \right\|_2, \\ i = 1, \ldots C, m = 1, \ldots, M, \end{aligned} \tag{8}$$

where ω_m is the regulating coefficient on the m-th modality and can be obtained by learning or use empirical value for simplicity.

3. Experiments

In this section, we evaluate our method HSRC on real world data and perform comparisons with some related methods. Experiments on data of single modality and multiple modalities are both conducted.

3.1. Multiple feature description

The data used in experiments is of multi-view features extracted from Multiple Feature Data Set[1] (MF dataset). This dataset consists of 2000 samples of handwritten digits 0–9. For each digit, there are 200 samples with six different views, namely Fourier coefficients of the character shapes, profile correlations, Karhunen-Love expansion coefficients, pixel averages, Zernike moments and morphological characteristics. The descriptions of these features are shown in Table 1.

[1] 1 http://archive.ics.uci.edu/ml

3.2. Experimental setting and evaluation criteria

To evaluate our proposed HSRC method, we compare it with several state-of-the-art methods including linear SVM [18], Multi-Kernel SVM (MKL-SVM) [19], SRC [6], and Canonical Correlation Analysis [20] based SVM (denoted as CCA-SVM in the following context). These methods are widely used in classification tasks. Specifically for multi-modal data, CCA can make use of two views of the same semantic object to extract the representation of the semantics by correlating linear relationships between two multi-dimensional variables [21]; on the other hand, Multi-Kernel SVM uses inputs coming from different representations (kernels) from different views and then combines the multiple kernels through a weighted sum.

Given the MF dataset with 10 digits, we distinguish two different digits (such as “6”and “9”) from the selected samples subset, where each digit subset include not only the samples of the current digit but also other digit samples (we called noisy samples). For our experiments, totally 420 samples are selected containing 400 handwritten digit samples of “6” and “9” and 20 noisy samples. Classifiers for features of single view and multiple views are applied separately.

To quantitatively evaluate the classification performance, some quantitative measurements including Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Mean Predictive Value (MPV) are employed. Sensitivity measures the proportion of positives that are correctly identified, and specificity measures the proportion of negatives that are correctly identified. PPV and NPV are respectively the proportions of positive and negative results in tests that are true positive and true negative. MPV is the average values of accuracy. Formulas for these measurements are given below.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}, \quad SEN = \frac{TP}{TP+TN}, \quad SPE = \frac{TN}{TN+FP},$$
$$PPV = \frac{TP}{TP+FP}, \quad NPV = \frac{TN}{TN+FN}, \quad MPV = \frac{SEN+SPE}{2}.$$

Here *TP*, *TN*, *FP*, *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively.

Classification performance is evaluated with 10-fold cross-validation. For each round of training, if not explicitly mentioned, classifiers are constructed by randomly selecting 90% of the training samples as training data and using the rest as testing data to measure the classification performance. We repeat the experiments 10 times for avoiding bias. The best parameters are determined through an inner 5-fold cross-validation on the training stage. We obtain the optimal values of λ_h, $h = 1, \ldots, H$ for our HSRC method by grid search in the range of $\{10^{-3}, 10^{-2}, 10^{-1}\}$.

3.3. Experimental results on MF handwritten dataset

In our experiments, we use 420 samples from the multiple features handwritten digits dataset, consisting of 200 digit “6”, 200 digit “9” and 20 other digits with six views (mfeat-fou, mfeat-fac, mfeat-kar, mfeat-pix, mfeat-zer and mfeat-mor). Each sample has 76, 216, 64, 240, 47 and 6 elements for these six views, respectively. A dictionary with 2 layers (H=2) is constructed in this study.

3.3.1. Classification results on single modality

For single modality, we compare our method with two above-mentioned methods (SVM, SRC) for classification between digit “6” and “9”. The classification results with quantitative measurements are listed in Table 2 for six different feature views.

Table 2
Classification results of the proposed HSRC method and baseline methods.

View	Method	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)
mfeat-fac	SVM	98.2	98.1	98.4	98.5	98.2
	SRC	96.6	96.5	96.5	96.9	96.3
	HSRC	**98.7**	**98.5**	**98.9**	**98.9**	**98.5**
mfeat-mor	SVM	63.8	68.8	58.8	62.4	68.9
	SRC	60.3	57.0	**63.6**	59.5	64.0
	HSRC	**65.3**	**69.4**	61.3	**64.1**	**70.5**
mfeat-fou	SVM	66.4	72.5	60.3	64.9	71.7
	SRC	66.8	72.6	61.1	65.6	72.1
	HSRC	**69.3**	**76.3**	**62.4**	**67.8**	**74.3**
mfeat-kar	SVM	98.2	**98.0**	98.2	98.3	**98.1**
	SRC	98.2	97.8	98.6	98.6	97.9
	HSRC	**98.3**	97.8	**98.9**	**98.9**	97.9
mfeat-pix	SVM	98.5	**98.1**	98.8	98.8	**98.2**
	SRC	98.2	97.4	99.0	99.1	97.5
	HSRC	**98.8**	97.9	**99.7**	**99.7**	98,0
mfeat-zer	SVM	64.8	68.3	**61.4**	63.8	69.7
	SRC	64.0	71.6	56.4	62.8	70.6
	HSRC	**66.8**	**72.8**	60.8	**65.5**	**72.1**

As can be seen from Table 2, HSRC exhibits higher performance for six different views compared with SVM and conventional SRC. However, the classification accuracies with all three methods for feature views of mfeat-fou, mfeat-mor and mfeat-zer are unsatisfactory. The underlying reason is that these features alone cannot supply sufficiently discriminatory information for classification. But as shown in next section, they still can contribute to enhancing the classification performance by combining with other features.

3.3.2. Classification results on multiple modalities

Multiple modalities data can provide complementary information to improve classification performance. We compare the performance of the competing methods including MKL-SVM, CCA-SVM, conventional SRC and our proposed HSRC to distinguish the handwritten digits with different combinations of two feature views. For the sake of fairness, we use the same modal fusion strategy for the conventional SRC and our proposed HSRC method shows the classification results of handwritten digits for 15 different feature combinations, where our HSRC method obtains the best classification accuracies among the listed competing methods.

3.5. Discussion

Obviously, the set of selected features under this setting would be appropriate if we are planning to build a linear classification modal (e.g. linear SVM). This is because these features are selected to minimize redundancy and maximize relevance to the class labels in the original feature space.

For mfeat-mor, mfeat-fou and mfeat-zer views, respectively, the classification performance of two different digits is not good for all the competing methods in Table 2. We use the complementary information of different views to re-evaluate the classification performance, from Table 3 we can see, combining different complementary information from multiple views can definitely improve the classification performance.

In this work, we just take a simple feature fusion measure to combine the multiple modalities. In the future, we can consider the more effective fusion method and emerge the multiple modalities into the construction of deep dictionary. As for the layer number H of the deep dictionary, its optimal value should be chosen by learning instead of taking a fixed number (we simply let H equal 2 in experiments).

4. Conclusion

In this paper, we propose to enhance the conventional sparse representation based classification by augmenting the single-layer dictionary to the deep dictionary. Thus, our proposed hierarchical

Table 3
Classification accuracies of the competing methods for handwritten digits.

View1	View2	MKL-SVM (%)	CCA-SVM (%)	SRC (%)	HSRC (%)
mfeat-fac	mfeat-mor	98.4	98.0	98.1	**98.8**
mfeat-fac	mfeat-fou	98.3	98.2	98.2	**98.7**
mfeat-fac	mfeat-kar	98.3	98.3	98.0	**99.0**
mfeat-fac	mfeat-pix	98.3	98.2	98.1	**99.3**
mfeat-fac	mfeat-zer	98.5	98.4	98.0	**98.8**
mfeat-mor	mfeat-fou	74.5	73.2	72.7	**75.3**
mfeat-mor	mfeat-kar	98.2	98.0	98.1	**98.5**
mfeat-mor	mfeat-pix	99.0	98.5	98.4	**98.9**
mfeat-mor	mfeat-zer	76.2	75.1	75.5	**77.5**
mfeat-fou	mfeat-kar	98.1	98.1	98.2	**98.4**
mfeat-fou	mfeat-pix	98.6	98.5	98.3	**98.9**
mfeat-fou	mfeat-zer	75.5	75.4	75.4	**76.7**
mfeat-kar	mfeat-pix	98.5	98.4	98.2	**98.8**
mfeat-kar	mfeat-zer	98.3	98.2	98.2	**98.4**
mfeat-pix	mfeat-zer	98.6	98.5	98.3	**98.8**

sparse representation based classification method can overcome some issues through the use of more efficient layer-by-layer representation. Experiments on the classification of handwritten digits, especially with multi-view data, show that our method can achieve more accurate classification results compared with the state-of-the-art counterpart classification methods. It may reveals that our propose HSRC method can better take advantage of the complementary information from multiple modalities to improve the classification accuracy and robustness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61273021, 61403281, 61471225), Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1500501, KJ1600512, KJ1600518, KJ100415), Natural Science Foundation of Chongqing (cstc2014jcyjA30013, cstc2015jcyjA70010), Natural Science Foundation of Hainan (20156219), Natural Science Foundation of Shandong (ZR2014FM002), and a China Postdoctoral Science Special Foundation Funded Project (2015T80717).

References

[1] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Netw. Learn. Syst. (2016).

[2] F. Liu, C.-Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification, NeuroImage 84 (2014) 466–475.

[3] Y. Jin, Y. Shi, L. Zhan, P.M. Thompson, Automated multi-atlas labeling of the fornix and its integrity in Alzheimer's disease, in: IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 140–143.

[4] Z. Wang, X. Zhu, E. Adeli, Y. Zhu, C. Zu, F. Nie, D. Shen, G. Wu, Progressive Graph-Based Transductive Learning for Multi-modal Classification of Brain Disorder Disease, MICCAI, Springer, 2016, pp. 291–299.

[5] 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 (2009) 210–227.

[6] K. Huang, S. Aviyente, Sparse representation for signal classification, Adv. Neural Inf. Process. Syst. 19 (2006) 609–616.

[7] X. Zhu, H.-I. Suk, D. Shen, Sparse discriminative feature selection for multi–class alzheimer's disease classification, Machine Learning in Medical Imaging, Springer, 2014, pp. 157–164.

[8] M. Yang, L. Zhang, J. Yang, D. Zhang, Regularized robust coding for face recognition, IEEE Trans. Image Process. 22 (2013) 1753–1766.

[9] L.N. Tan, A. Alwan, G. Kossan, M.L. Cody, C.E. Taylor, Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data, J. Acoust. Soc. Am. 137 (2015) 1069–1080.

[10] C. Li, Y. Ma, X. Mei, C. Liu, J. Ma, Hyperspectral image classification with robust sparse representation, IEEE Geoscience and Remote Sensing Letters 13 (2016) 641–645.

[11] W. Deng, J. Hu, J. Guo, Extended SRC: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 1864–1870.

[12] W. Deng, J. Hu, J. Guo, In defense of sparsity based face recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013), pp. 399–406.

[13] L. Ma, C. Wang, B. Xiao, W. Zhou, Sparse representation for face recognition based on discriminative low-rank dictionary learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2586–2593.

[14] X. Jiang, J. Lai, Sparse and dense hybrid representation via dictionary decomposition for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1067–1079.

[15] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.

[16] D. Bertsekas, Constrained optimization and Lagrange multiplier methods, Academic press, 1996.

[17] X. Zhu, H.-I. Suk, D. Shen, Multi-modality canonical feature selection for Alzheimer's disease diagnosis, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer, 2014, pp. 162–169.

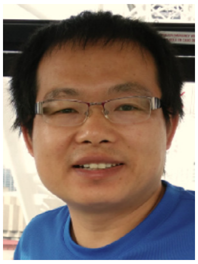
[18] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.

[19] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (2011) 2211–2268.

[20] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.

[21] B. Thompson, Canonical correlation analysis, Encyclopedia of Statistics In Behavioral Science, John Wiley & Sons Ltd, 2005.

Zhengxia Wang received her PH.D. degree in computer software and theory from Chongqing University, China. She is currently an associate professor at Chongqing jiaotong University. Her main research interests include genetic regulatory network, stability of dynamical system and application in image processing. Her current research interests mainly focus on machine learning, pattern recognition and image processing.



Shenghua Teng received the B.S. degree from North China Electric Power University in 2000, M.S. degree from Beijing University of Aeronautics and Astronautics in 2003, and Ph.D. degree from Institute of Electronics, Chinese Academy of Sciences in 2006. Currently he is with Shandong University of Science and Technology as an associate professor. His main research interests are image processing and pattern recognition.

Guodong Liu received his Master Degree in geographic information science from Taiyuan University of Technology, China. His main research interests include application of 3S in transportation industry. His current research interests mainly focus on spatial analysis, spatial statistics and image processing.

Zengshun Zhao received the Ph.D. degree in control engineering from the Institute of Automation, Chinese Academy of Sciences, in 2007. In 2011, he worked as a visiting scientist with Prof. C.S. Zhang at Tsinghua University. From 2012 to 2014, he worked as a postdoctoral at College of Control Science and Engineering, Shandong University. He is currently an associate professor at the College of Electronics, Communication and Physics, Shandong University of Science and Technology, Qingdao, China. His research interests include intelligent robot, machine learning & pattern recognition, Computer vision, and the computational intelligence.

Hongli Wu graduated from Chinese Academy of Sciences, Chengdu computer application Institute in 2010 and received a Doctorate degree. Her research direction is the theory of software engineering, software process technology and method, and data mining.