Joint Feature Selection with Low-rank Dictionary Learning

Homa Foroughi
homa@ualberta.ca
Moein Shakeri
shakeri@ualberta.ca
Nilanjan Ray
nray1@ualberta.ca
Hong Zhang
hzhang@ualberta.ca

Department of Computing Science
University of Alberta
Edmonton, Canada

Abstract

Feature selection is one of the well known dimensionality reduction methods that efficiently describes the input data by removing irrelevant variables and reduces the effects of noise to provide good prediction results. In this paper, we propose a feature selection method by integrating dictionary learning and low-rank matrix approximation and apply it to image classification. The objective function finds a subset of features by preserving the reconstructive relationship of the data. This is achieved by minimizing the within-class reconstruction residual and simultaneously maximizing the between-class reconstruction residual. Simultaneously, the $l_{2,1}$-norm minimization on projection matrix is applied to jointly select the most relevant and discriminative features. The combination of low-rank approximation and Fisher discrimination dictionary learning, leads in more compactness within the same class and dissimilarity between different classes. As a result, even a simple classifier like KNN would perform surprisingly well and classify data accurately. Our proposed method is extensively evaluated on different benchmark image datasets and shows superior performance over several feature selection methods. The experimental results together with the theoretical analysis validate the effectiveness of our method for feature selection, and its efficacy for image classification.

1 Introduction

In many areas, such as computer vision and pattern recognition, data are characterized by high dimensional feature vectors. If these vectors are processed directly, it usually leads to difficult pattern recognition task because of the curse of dimensionality. However, in practice, only a small subset of features is really important and discriminative. So, for the efficient processing of a high dimensional feature, its dimensionality has to be reduced without a loss in the original properties. In literature, there are mainly two distinct ways for dimensionality reduction: feature extraction and feature selection (FS). Feature extraction transforms features form high dimensional patterns space into a lower space by combining several original features, while FS chooses a subset of features by eliminating the irrelevant

Pages 97.1-97.13
DOI: https://dx.doi.org/10.5244/C.29.97

and redundant features based on certain criteria [28]. The fundamental part of FS is to determine a minimal feature subset, which can efficiently describe the input data. This problem is essentially a combinatorial optimization problem which is computationally expensive. Most of the traditional FS algorithms address this issue by evaluating the importance of each feature individually and selecting top-ranked features one by one. However, these techniques fail to provide any guarantee of global optimality [28]. Consequently, the correlation among features is neglected [31]. As a solution, researchers introduced joint FS by taking into account the relationship of different features. For instance, Cai *et al*. [5] proposed a two-step FS algorithm by incorporating spectral regression and l_1-norm regularization.

Recently, sparsity regularization in dimensionality reduction has been widely investigated and also applied into FS studies. l_1-SVM was proposed to perform FS using the l_1-norm regularization that tends to give sparse solution [2]; that was further improved [26] by combining both l_1-norm and l_2-norm to form a more structured regularization. Nie *et al*. [17] proposed a FS method with emphasizing joint $l_{2,1}$-norm minimization on both loss function and regularization. Yang *et al*. [31] combined the manifold learning and $l_{2,1}$-norm minimization into joint FS and proposed an unsupervised FS algorithm. Most recently, Yan *et al*. [28] introduced the sparse representation-based classification (SRC) [27] measurement criterion into FS and designed a joint sparse discriminative FS (JSDFS) method.

The so-called JSDFS method [28] achieves impressive results compared to other FS methods. Based on the assumption of SRC, their method selects a subset of features which minimize the within-class reconstruction residual and simultaneously maximize the between-class reconstruction residual in the subset of selected features. Nonetheless, the complexity of SRC can be very high due to using all the training samples and the discriminative information in the training samples is not sufficiently exploited by such a naive method [30]. As a result, the reconstructive relationship of samples could not be persevered well and the selected features are not discriminant enough. To overcome the drawbacks associated with the SRC algorithm, in this paper we propose a new FS method by learning a smaller-sized dictionary from the given training images while maintaining the sparse reconstruction relationship among samples. The main contributions of this paper are:

- Our FS method integrates dictionary learning (DL) and low-rank (LR) approximation to preserve the reconstructive relationship of data. Specifically, we formulate FS problem under LR dictionary learning with Fisher discrimination regularization.
- Optimizing the training samples from each class to be LR, reduces the diversity across items within each class which may affect the representation power of dictionary. However, incorporating Fisher discrimination on both class-specific representations and sparse coefficients provide enough discriminating ability into our framework. The representative bases learned by the proposed method are encouraged to be close within the same class, and far between different classes. Hence, we can achieve good accuracy on image classification, even with a simple classifier such as KNN or SVM.
- Our proposed method can preserve the intraclass compactness and interclass separability much better, compared to similar methods. As a result, the useful features can be preserved, while the irrelevant ones would be discarded during the projection process.

Extensive experimental results together with the theoretical analysis validate the effectiveness of our method for FS, and its feasibility of being applied to image classification task. The remainder of this paper is organized as follows. We first present some background in section 2. The proposed FS method is introduced in section 3. The experimental results on

benchmark datasets are conducted in section 4, followed by discussions. Finally, we provide concluding remarks in section 5.

2 Background

Recently, Yan *et al*. [28] introduced SRC-based measurement criterion into FS and designed a joint sparse discriminative FS method. Considering the decision rule of SRC, their objective function aims to find a subset of features, whose components could be well approximated by the linear combination of other components in the same class and this is achieved by minimizing the ratio of within-class reconstruction residual to between-class reconstruction residual in the subset of selected features. Although they can achieve promising results compared to other FS methods, it is well-known that SRC suffers from major drawbacks such as high computational complexity and low discriminativity of sparse coefficients. Also, noise and trivial information can make it ineffective [30]. More importantly, since the sparse coefficients and naive dictionary are not discriminative, the reconstruction scatter matrices would not preserve the reconstructive relationship of data well. Hence, the projection matrix obtained by minimizing their ratio is not optimal either, which means the selected features are not discriminant enough. These problems can be addressed by learning properly a dictionary from training samples. The sparse coefficients and the dictionary obtained by supervised DL methods are significantly more discriminative compared to those of SRC [30].

In the recent years, DL for sparse representation has attracted much attention and has been successfully applied to a variety of computer vision tasks. Unsupervised DL methods, do not utilize class information of training samples and their goal is to minimize the reconstruction error [29]. Although these methods can achieve promising results in image restoration [1], they are not advantageous for image classification. With the class labels of training samples available, the supervised DL methods exploit the class discrimination information, which results in better classification performance [29]. Most of the supervised DL methods learn such an adaptive dictionary mainly in two ways: either directly forcing the dictionary, or the sparse coefficients to be discriminative (usually through simultaneously learning a classifier) to promote the discrimination power of the dictionary [11]. In the former group, multiple or category-specific dictionaries are learned to promote discrimination between classes. In contrast, in the latter, discrimination is achieved by incorporating discriminative terms such as linear predictive classification error [32] and label consistency constraint [11] into the objective function.

Different from the most of class-specific DL methods, Yang *et al*. [30] introduced Fisher discrimination both in the sparse coding coefficients and class-specific representations which would further enhance the discrimination of the dictionary. The discrimination capability of their method, Fisher discrimination dictionary learning (FDDL) originates from two facts. First, each sub-dictionary is trained to have good representation power to the samples from the corresponding class, but have poor representation power to the samples from other classes. Second, the sparse coefficients are made discriminative through minimizing the within-class scatter and maximizing the between-class scatter of them. Both of these properties make FDDL a good choice for finding the discriminative sparse coefficients of training samples. Nevertheless, all the DL algorithms, including FDDL, work well when the input images are clean or corrupted by small noise. The performance of these methods deteriorates when the training data is contaminated, e.g., because of occlusion, disguise, lighting variations or pixel corruption [33].

LR matrix recovery, which determines a LR data matrix from corrupted data, has been successfully applied to different tasks including image classification. Inspired by [15, 33] which use LR matrix recovery to improve the performance of DL algorithm with noises, we integrate rank minimization into sparse representation for DL. The introduction of Fisher discrimination into the LR matrix recovery, makes the sub-dictionaries as independent as possible and promotes the discrimination power of sub-dictionaries toward each other. Consequently, the reconstruction scatter matrices would preserve the reconstructive relationship of data much better. Based on above discussion, in this work, we propose a Joint Feature Selection method using Low-rank Dictionary Learning (JFS-LDL) in order to capture discriminative features well.

3 Joint FS using Low-rank Dictionary Learning

Given a set of training data vectors $X = [X_1, X_2, \ldots, X_K] \in R^{m \times N}$, where X_i is the samples from i^{th} class, m is the feature dimension, and N is the total number of training samples.

3.1 Low-rank Approximation

The samples in class i are linearly correlated in many situations. More precisely, the matrix $X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,K_i}]$ should be approximately low-rank. LR matrix recovery seeks to decompose a data matrix X into $L+E$ by minimizing the rank of matrix L, while reducing $\|E\|_0$, the associated sparse noise. Since the aforementioned optimization problem is NP-hard, Cande's *et al.* [6] solve the following formulation to make the original LR tractable:

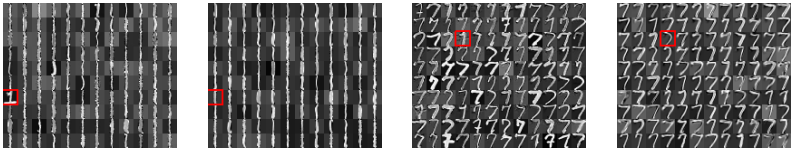
$$\min_{L,E} \quad \|L\|_* + \lambda \|E\|_1 \qquad s.t. \quad X = L + E \tag{1}$$

Using (1), it is possible to find the LR and sparse noise of samples of the i^{th} class; i.e., $X_i = L_i + E_i$. This step reveals the structural information of each class and makes training samples of that class more correlated. To solve the optimization problem of (1) efficiently, the technique of inexact augmented Lagrange multipliers (ALM) [6, 13] is usually applied due to its computational efficiency.

3.2 Low-rank Dictionary Learning using Fisher Discrimination

To improve the performance of FDDL with noises and contamination, we use a more robust representation of X_i, which is basically its LR representation, L_i in the FDDL objective function. Furthermore, when the standard LR matrix recovery is combined with Fisher discrimination, the images tend to be more similar to each other for the same class, which means more compactness exists within the same class and dissimilarity between different classes. Therefore, the learned sub-dictionary D_i would have better discrimination and reconstruction capabilities compared to FDDL model. Accordingly, the quality of structured dictionary $D = [D_1, D_2, \ldots, D_K]$ will influence the discriminativeness of the sparse coefficients A.

Denote by $L = [L_1, L_2, \ldots, L_K]$ the structured LR representations of training samples. D should have the capability to represent the sparse coefficients, i.e., $L \approx DA$. We can write A, the sparse coefficients of L over D, as $A = [A_1, A_2, \ldots, A_K]$, where A_i is the representation matrix of L_i over D. Following FDDL notations, A_i can be written as $A_i = [A_i^1; \ldots; A_i^j; \ldots; A_i^K]$



(a) (b) (c) (d)

Figure 1: Learned atoms of digits 1 and 7 of USPS dataset using (a),(c) original data vectors and (b),(d) their LR representation as the input of DL model

where A_i^j is the representation coefficients of L_i over D_j. The objective function of LR dictionary learning using Fisher discrimination is formulated as:

$$\min_{D,A} \sum_{i=1}^{K} \left(\|L_i - DA_i\|_F^2 + \|L_i - D_i A_i^i\|_F^2 \right) + \lambda_1 \|A\|_1$$
$$+\lambda_2 \left(tr(S_W(A) - S_B(A)) + \eta \|A\|_F^2 \right) \quad s.t. \quad \|d_n\|_2 = 1 \ \ \forall n; \quad \|D_j A_i^j\|_F^2 \leq \varepsilon_f, \ \ \forall i \neq j \quad (2)$$

where $S_W(A)$ and $S_B(A)$ are within-class and between-class scatter matrices of sparse coefficients A which are defined as:

$$S_W(A) = \sum_{i=1}^{K} \sum_{a_k \in A_i} (a_k - m_i)(a_k - m_i)^T \quad \text{and} \quad S_B(A) = \sum_{i=1}^{K} n_i (m_i - m)(m_i - m)^T \quad (3)$$

where m_i and m are the mean vectors of A_i and A, respectively. The objective function (2) can be further simplified and eventually divided into two sub-problems by optimizing D and A alternatively, i.e., updating A with D fixed, and updating D with A fixed. The optimization details can be found in [30]. Figure 1 compares sub-dictionaries learned by FDDL model (X_i as input of model) and LR-FDDL (objective function (2), L_i as input of model) on two digits from USPS dataset. As can be seen, the variations in the shape, thickness and orientation have been significantly removed by sparse noise. Clearly, LR reduces the diversity across items within each class and consequently dissimilarity between different classes would be increased, which means sub-dictionaries are more discriminant toward each other. Rate of misclassification of similar images from different classes (e.g. 1 and 7 in red squares) would be decreased consequently. More importantly, due to dissimilarity of classes, even a simple classifier like KNN can perform surprisingly well for the classification task.

3.3 Joint Feature Selection

We aim to select a subset of features that preserve the sparse reconstructive relationship of the training samples. This is achieved by minimizing the within-class reconstruction residual error and simultaneously maximizing the between-class reconstruction residual in the subset of selected features. Based on the discussion in section 2, to promote discrimination, we exploit the sparse coefficients obtained by objective function (2) to obtain the within-class (S_W^L) and between-class (S_B^L) scatter matrices:

$$S_W^L = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[L_{i,j} - D_i A_{i,j}^i \right] \left[L_{i,j} - D_i A_{i,j}^i \right]^T$$
$$S_B^L = \frac{1}{N(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \sum_{\substack{s=1 \\ s \neq i}}^{K} \left[L_{i,j} - D_s A_{i,j}^s \right] \left[L_{i,j} - D_s A_{i,j}^s \right]^T \quad (4)$$

Algorithm 1 JFS-LDL Algorithm

Input: Data matrix X
Output: Projection matrix P
1: Find LR representation of $X_i, \forall i = 1,\dots,K$ by Eq. 1
2: Find D and A by LR dictionary learning using Fisher discrimination by Eq. 2
3: Construct scatter matrices S_W^L and S_B^L using Eq. 4
4: Solve the eigen-problem, find largest eigen-vectors to form Y
5: **Initialize:**parameters $\beta, M, \mu, max_\mu, \rho$
6: **while** $\left\|D^T P + E - Y\right\|_\infty < \varepsilon$ **do**
7: Update E as: $E = \frac{1}{2+\mu}(-M + \mu Y - \mu D^T P)$
8: Update P as: $P = (2\beta S + \mu D D^T)^{-1}(\mu D Y - D M - \mu D E)$
9: Update M as: $M = M + \mu(D^T P + E - Y)$
10: Update μ as: $\mu = min(\rho\mu, max_\mu)$

where n_i is the number of training samples in i^{th} class, $L_{i,j}$ is the LR representation of the $X_{i,j}$, the j^{th} training sample in class i. Denote by A_i^i the sparse representation of L_i over D_i; $A_{i,j}^i$ implies its j^{th} column. Similarly, $A_{i,j}^s$ is the sparse coefficients of $L_{i,j}$ over D_s. To optimally preserve the sparse reconstructive relationship of data and simultaneously achieving row-sparsity, the projection matrix $P \in R^{m\times m}$ is found by the following optimization problem:

$$\min_P \quad \frac{tr(P^T S_W^L P)}{tr(P^T S_B^L P)} + \beta \left\|P\right\|_{2,1} \tag{5}$$

That is to say, the new representation of image x after FS is obtained as; $x' = P^T x$ where $x'(k) = x(k)$ if the k-th feature is selected; otherwise $x'(k) = 0$. The projection matrix obtained by (5), can preserve within-class compactness and between-class separability well enough in the low-dimensional space. As a result, the projected samples are more discriminative and simultaneously retain the intrinsic properties of data. By imposing $l_{2,1}$-norm constraint in the objective function, P is simultaneously optimized to achieve row-sparsity; consequently, the useful features can be preserved, while the irrelevant features can be discarded [22]. According to the ratio trace problem [7], (5) can be reformulated as:

$$\min_P \quad tr(P^T S_W^L P) + \beta \left\|P\right\|_{2,1} \quad s.t. \quad tr(P^T S_B^L P) = I \tag{6}$$

Although the objective function (6) is convex, the constraint is not. Based on the theorem 1 of [4], P can be obtained through the following two steps:

1. Solve the eigen-problem $S_W^L Y = \Lambda S_B^L Y$ to find Y
2. Find P which satisfies $D^T P = Y$

where Y is the matrix of generalized eigenvectors corresponding to $min(N,m)$ largest eigenvalues, Λ is a diagonal matrix whose diagonal elements are eigenvalues, and D is the structured dictionary. Finding a solution for P under $l_{2,1}$ constraint such that $D^T P = Y$ is usually impossible. Therefore, a residue matrix E is introduced [28] and the following problem is solved instead:

$$\min_{P,E} \quad \left\|E\right\|_F^2 + \beta \left\|P\right\|_{2,1} \quad s.t. \quad D^T P + E = Y \tag{7}$$

For solving (7) efficiently, we take a similar approach to [28], which is basically an iterative algorithm based on inexact ALM [13]. The augmented Lagrangian function of (7) is defined

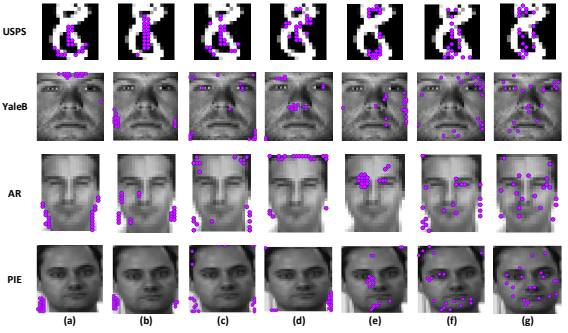


Figure 2: Twenty selected features by (a)MIM (b)LS (c)mRMR (d)Relief (e)SDFS (f)JSDFS and (g)JFS-LDL methods on four datasets

as follows:

$$L(P,E,M,\mu) = \|E\|_F^2 + \beta\|P\|_{2,1} + \frac{\mu}{2}\|D^T P + E - Y\|_F^2 + \langle M, D^T P + E - Y\rangle \qquad (8)$$

where M is the Lagrange multipliers and μ is a positive parameter. ALM method alternatively updates the variables P and E by iteratively minimizing the augmented Lagrangian function. Algorithm 1 outlines the details of solving (8). In this algorithm, S is a diagonal matrix $S_{ii} = 1/2\|P_i\|_2$. We set $\beta = 10^{-6} \times \|\bar{D}\|_F^2$ and the parameters M, μ, max_μ and ρ are set to zero matrix, 10^{-6}, 1.01 and 2, respectively. Details of the intermediate steps can be followed in [28].

3.4 Training Time Complexity

- To find LR representation of X_i for all K classes, we use the accelerated version of robust orthonormal subspace learning (ROSL+) [23] which its complexity is bounded by $O\big(r^2(m+n)\big)$, where r is the rank of matrix L. So, the complexity of this step would be $O\big(Kr^2(m+n)\big)$.
- The complexity of LR dictionary learning using Fisher discrimination to find D and A would be similar to that of FDDL, consisting of updating sparse coefficients and sub-dictionaries. The overall time complexity of simplified version, which we used, is $t\big(\sum_i n_i O(m^2 p_i^\varepsilon) + \sum_i p_i O(mn_i)\big)$, where t is the total number of iterations of this step, p_i is the number of i^{th} sub-dictionary atoms and $\varepsilon \geq 1.2$ is a constant [30].
- The time complexity of constructing scatter matrices S_W^L and S_B^L is $\sum_i n_i O(m^2 + mp_i)$ and $\sum_i n_i O\big(K(m^2 + mp_s)\big)$ respectively, where $s \neq i$.
- For eigen-decomposition in order to find Y, we exploit Lanczos algorithm [18] to compute the top $d = min(N,m)$ eigenvectors with $O(dm^2)$.
- Finally, to find P we utilize inexact ALM as shown in the steps 6-10, which its complexity is $O\big(\gamma(m^3 + 6mP^2)\big)$, where $P = \sum_i p_i$ and γ is the number of iterations.

4 Experimental Results

We conduct extensive experiments on several datasets to verify the effectiveness of the proposed JFS-LDL method in comparison with other FS methods and validate its capability for

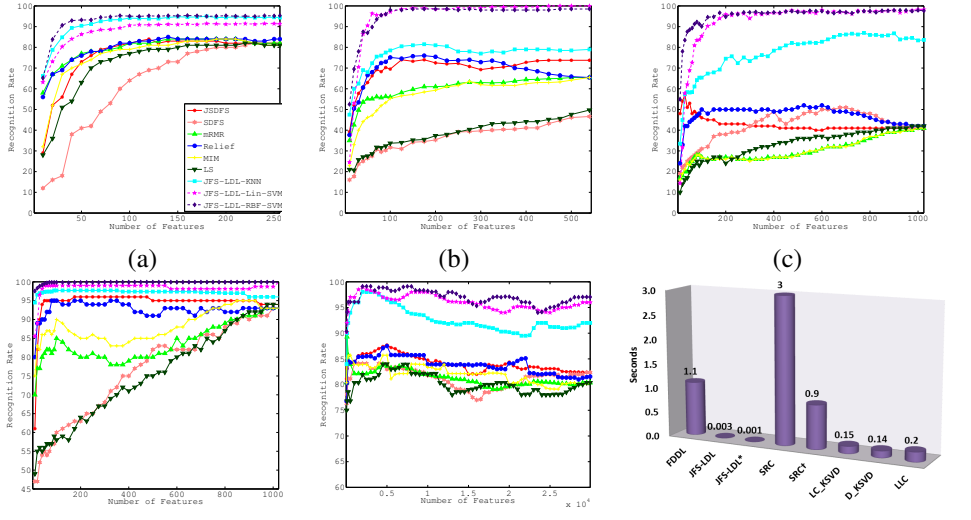


Figure 3: Recognition rate vs. number of selected features on five datasets (a)USPS (b)AR (c)YaleB (d)PIE (e)UCF using various FS methods. The recognition rate of JFD-LDL with KNN, Lin-SVM and RBF-SVM are shown by solid squares, dashed pentagrams and dashed diagonals respectively. (f)Average classification time of an image in USPS dataset.

image classification task.

(a) **Face Recognition**: We use three benchmark datasets; the Extended YaleB [8] contains 2,414 frontal face images of 38 human subjects under different illumination conditions. All the face images are cropped and resized to 32×32 and we randomly select 32 images per class for training and the rest for test. Another dataset is the CMU PIE [24] which contains 68 individuals with 41,368 face images including different poses, illumination conditions and expressions. In our experiments, we just use the near frontal pose (C27) which leaves us about 100 face images for each individual, 30 of which are randomly chosen for training and the rest is used for test. All Images are resized to 32×32. Finally, we evaluate our method on the AR face dataset [16] that consists of over 4,000 frontal images from 126 individuals. As a standard evaluation procedure, we select a subset of 2,600 images from 50 male and 50 female subjects in the experiments. Each face image is resized to 27×20 and for each person, we randomly select 20 images for training and the remaining disguise images with scarves or sunglasses are used for testing.

(b) **Digit Recognition**: We also perform handwritten digit recognition on the widely used USPS dataset [10], which has 7,291 training and 2,007 test images, each of size 16×16. Here, the number of atoms in each sub-dictionary is set to 200.

(c) **Action Recognition**: At last, we conduct action recognition on the UCF sport action dataset [20]. There are 140 videos which cover ten sport action classes such as diving, golfing, kicking, lifting and so on. Their action bank features can be found in [21], which has around 30,000 feature dimensions. We follow the experiment settings in [11, 30] and evaluate our method via fivefold cross validation.

	SRC [27]	SRC† [27]	FDDL [30]	LC-KSVD [11]	LLC [25]	D-KSVD [32]	JFS-LDL	JFS-LDL⋆
USPS	93.9	78.5	97.1	96.4	95.5	68.9	95.3 (**160**)	90.1
YaleB	97.2	80.2	97.0	96.7	90.7	94.1	97.8 (**500**)	94.5
AR	97.5	66.5	92.7	93.7	88.7	88.8	98.5 (**150**)	90.2
PIE	93.0	90.2	97.0	91.8	90.3	89.3	99.9 (**300**)	99.7
UCF	92.9	83.6	94.3	91.2	87.5	88.1	99.1 (**2000**)	99.0

Table 1: The recognition accuracy (%) of different methods on various datasets

First, we compare the proposed method with several standard FS methods including minimum Redundancy Maximum Relevance(mRMR) [19], Laplacian Score(LS) [9], Mutual Information Maximization(MIM) [8], Relief [12] and two sparsity-based methods proposed in [28]; sparse discriminative feature selection (SDFS) and joint sparse discriminative feature selection (JSDFS). Figure 2 shows 20 selected features on some training images from four different datasets. We observe that the selected features by MIM, mRMR, LS and SDFS have concentrated distribution, while those by JSDFS and JFS-LDL are distributed dispersedly and this is mainly due to joint FS property of $l_{2,1}$-norm. Moreover, compared with JSDFS and Relief, the selected features by our method are distributed in areas that hold more discriminative information; e.g., in the face datasets, these points are mostly around eyes, nose and mouth.

Then, we evaluate our method on various image classification tasks with different number of selected features. KNN is used for classification on the new representations of images. Figure 3 illustrates the recognition rate vs. the number of selected features on five datasets. As these graphs illustrate, JFS-LDL (solid line with squares) improves the recognition rate over other methods consistently and notably. JFS-LDL can maintain a relatively stable performance under different dimensions, and as the number of selected features decreases, its advantage becomes more obvious. We further evaluate JFS-LDL based on another popular classifier, SVM with linear and RBF kernels. We use One-Against-All SVM for multi-class classification and the parameters are selected by cross-validation. These results are also included in Figure 3 with dashed pentagrams and dashed diagonals for linear and RBF SVM respectively. Both linear and RBF SVM tend to have higher accuracy than KNN constantly across all dimensions in each dataset; however, in some datasets such as YaleB, they affect the recognition performance significantly. Generally, SVM shows more stable trend than KNN especially in lower dimensions.

Next, we compare the performance of our method in image classification to that of SRC and some of the recently proposed DL methods. The detailed comparison results on five datasets are summarized in Table 1. Here, we choose SVM with RBF kernel for classification of JFS-LDL and report the best performance as well as the corresponding number of selected features in parentheses. Additionally, we measure the performance of SRC when using the same size as the dictionary (denoted SRC†) and that of our method while using just 10% of randomly sampled features (denoted JFS-LDL⋆). We observe that JFS-LDL with a selected subset of features is superior or competitive to other methods with much higher feature dimensions in all datasets. This implies the effectiveness of our method in capturing the discriminative information for classification. Moreover, it is noticed that the proposed method can achieve high recognition accuracy using a random subset of features, i.e., JFS-LDL⋆. Class-specific LR dictionary learning along with Fisher discrimination, encourages discrimination capability of dictionary; hence, we can achieve good accuracy on image classification, even with a simple classifier such as KNN or SVM. In contrasts to most of DL methods which use l_1-optimization to find the representation of test images and use the re-

	JFS-FDDL			JFS-LDL		
	KNN	Lin-SVM	RBF-SVM	KNN	Lin-SVM	RBF-SVM
USPS (**160**)	88.4	83.4	88.1	94.8	91.4	95.3
YaleB (**500**)	35.9	79.2	83.1	87.3	98.2	97.8
AR (**150**)	56.2	73.5	76.3	81.5	99.9	98.5
PIE (**300**)	80.8	91.5	94.6	99.9	99.9	99.9
UCF (**2000**)	81.2	94.1	95.6	98.0	99.0	99.1

Table 2: Performance comparison of JFS-LDL and JFS-FDDL on different datasets

construction error for classification, which is time-consuming, our classification schema is very efficient and fast. For instance, we compare the average classification time of an test image for the evaluated methods on USPS dataset in Figure 3(f). As can be seen, our method (using RBF-SVM) is much faster than SRC and other DL methods, which is a desirable property for large-scale image classification task.

Finally, to verify the efficacy of LR in the proposed FS method, we use the original FDDL model that uses raw data vectors X_i as the input of model and then perform JFS using Algorithm 1; clearly, step 1 would be ignored. Then, we do classification using KNN and SVM. Table 2 compares the recognition rate of JFS-LDL and JFS-FDDL on five datasets. It is noticed that JFS-LDL noticeably outperforms JFS-FDDL and this is mainly due to the ability of LR in denoising the data and thereby increasing the dissimilarity between classes. As a result, classes would be more separable and a simple classifier can achieve high recognition rate. The effect of LR is more noticeable in face datasets whose images are under sever occlusion, disguise and illumination variations. Using LR approximation gives us the opportunity to have well-separated classes and there is no need to use complicated and time-consuming classifiers like sparse coding (l_1-minimizer); however, it should be noted that the integration of LR approximation with Fisher discrimination would enable us to have enough discrimination to capture useful features.

5 Conclusion

In this paper, we proposed a joint FS method by integrating LR approximation and Fisher discrimination DL. The objective function finds a subset of features that preserve the reconstructive relationship of the data by minimizing the ratio of within-class reconstruction residual to between-class reconstruction residual in the subset of selected features. Meanwhile, the $l_{2,1}$-norm minimization on projection matrix is applied to jointly select features. As a result, the projected samples are more discriminative and simultaneously retain the important properties for classification, e.g., intraclass compactness and interclass separability, as well as the reconstructive relationship. Extensive experiments on benchmark datasets show that JFS-LDL consistently outperforms all the other evaluated FS methods, especially in lower dimensions in image classification task. Besides, the combination of LR approximation and Fisher discrimination, leads in more compactness within the same class and dissimilarity between different classes. Consequently, a simple and fast classifier like KNN or SVM would perform well for classification. The proposed method can achieve superior or competitive recognition rate compared to the recently proposed DL methods in much less time. The future work may include exploring the joint learning of discriminative features and dictionaries, which has been explored to some extent by Lu *et al*. [14] for image set based face recognition in order to have more discriminative information, that may be ignored in the feature learning stage if feature selection and dictionary learning are performed individually.

References

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[2] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[3] Gavin Brown. A new perspective for information theoretic feature selection. In *International conference on artificial intelligence and statistics*, pages 49–56, 2009.

[4] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression: A unified approach for sparse subspace learning. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 73–82. IEEE, 2007.

[5] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342. ACM, 2010.

[6] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[7] Keinosuke Fukunaga. Introduction to statistical pattern recognition. 1990. *Ch*, 9:401–405.

[8] Athinodoros S. Georghiades, Peter N. Belhumeur, and David Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[9] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

[10] Jonathan J. Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, 1994.

[11] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label consistent K-SVD: learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2651–2664, 2013.

[12] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256, 1992.

[13] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[14] Jiwen Lu, Gang Wang, Weihong Deng, and Pierre Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *Computer Vision–ECCV 2014*, pages 265–280. Springer, 2014.

[15] Long Ma, Chunheng Wang, Baihua Xiao, and Wen Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2586–2593. IEEE, 2012.

[16] A.M. Martinez and R. Benavente. The AR face database. Technical report, 1998.

[17] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint âĎŞ2,1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

[18] Beresford N Parlett. *The symmetric eigenvalue problem*, volume 7. SIAM, 1980.

[19] Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[20] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[21] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[22] Jun Shi, Zhiguo Jiang, Danpei Zhao, Hao Feng, and Chao Gao. Regularized least square discriminant projection and feature selection. *Journal of Electronic Imaging*, 23 (1):013003–013003, 2014.

[23] Xianbiao Shu, Fatih Porikli, and Narendra Ahuja. Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3874–3881. IEEE, 2014.

[24] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression database. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25 (12):1615–1618, 2003.

[25] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[26] Li Wang, Ji Zhu, and Hui Zou. Hybrid huberized support vector machines for microarray classification. In *Proceedings of the 24th international conference on Machine learning*, pages 983–990. ACM, 2007.

[27] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[28] Hui Yan and Jian Yang. Sparse discriminative feature selection. *Pattern Recognition*, 2014.

[29] Meng Yang, Dengxin Dai, Lilin Shen, and Luc Van Gool. Latent dictionary learning for sparse representation based classification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4138–4145. IEEE, 2014.

[30] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232, 2014.

[31] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2,1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1589. Citeseer, 2011.

[32] Qiang Zhang and Baoxin Li. Discriminative K-SVD for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[33] Yangmuzi Zhang, Zhuolin Jiang, and Larry S Davis. Learning structured low-rank representations for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 676–683. IEEE, 2013.