# The Role of Dictionary Learning on Sparse Representation-Based Classification

Soheil Shafiee, Farhad Kamangar and Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington, USA

## ABSTRACT

This paper analyzes the role of dictionary selection in Sparse Representation-based Classification (SRC). While SRC introduces interesting results in the field of classification, its performance is highly limited by the number of training samples to form the classification matrix. Different studies addressed this issue by using a more compact representation of the training data in order to achieve higher classification speed and accuracy. Representative selection methods which are analyzed in this paper include Metaface dictionary learning, Fisher Discriminative Dictionary Learning (FDDL), Sparse Modeling Representative Selection (SMRS), and random selection of the training samples. The first two methods build their own dictionaries via an optimization process while the other two methods select the representatives directly from the original training samples. These methods, along with the original method which uses all training samples to form the classification matrix, were examined on two face datasets and one digit dataset. The role of feature extraction was also studied using two dimensionality reduction methods, down-sampling and random projection. The results show that the FDDL method leads to the best classification accuracy followed by the SMRS method as the second best. On the other hand, the SMRS method requires a much smaller learning time which makes it more appropriate for dynamic situations where the dictionary is regularly updated with new samples. The accuracy of the Metaface dictionary learning method was specifically less than the other two methods. As expected, using all the training samples as the dictionary resulted in the best recognition rates in all the datasets but the classification times for this approach were far larger than the required time using any of the three dictionary learning methods.

## Categories and Subject Descriptors

G.1.6 [**Optimization**]: Constrained Optimization; I.5.4 [**Ap plications**]: Computer Vision

PETRA'13, May 29 - 31, 2013, Island of Rhodes, Greece.
Copyright ©2013 ACM ISBN 978-1-4503-1300-1/13/05... $15.00

## General Terms

Experimentation

## Keywords

Sparse Representation-based Classification, Metaface Learning, Fisher Discriminative Dictionary Learning, Sparse Modeling Representative Selection

## 1. INTRODUCTION

Classification is one of the most challenging problems in pattern recognition and it is widely used in different areas such as signal, image and speech processing. Recently, a novel classification approach, called Sparse Representation based Classification (SRC), is proposed by Wright et al. [21]. SRC has been used in different applications such as face recognition [21], signal classification [12] and activity recognition [22]. This classification method works based on the emerging theory of compressive sensing (CS) and sparse signal representation [6], [5]. In SRC, classification problem is mapped into an under-determined system of equations which is solved using $\ell^1$-norm minimization. The sparse solution determines the class of the unknown test sample. This linear system of equations can be considered as a dictionary problem where the original training samples are used to form its atoms. In [21], face recognition using this method is shown to achieve high recognition accuracy compared to other dimensionality reduction methods, such as Eigen faces [20], Fisher faces [4] and Laplacian faces [11]. The algorithm presented in [21] uses the entire training dataset to recognize each unknown test sample. This implies that the speed of the recognition task at run time is directly affected by the number of training samples. Given the high recognition accuracy of SRC, it becomes important to reduce the time and memory requirements of this method. Improvements in computational and memory efficiency of the SRC method makes it a more practical solution to be implemented for portable devices, and can also significantly decrease the computational load of SRC running on more powerful hardware. The time complexity of SRC is quadratic with respect to the number of training samples [7], and thus reducing the number of training samples by a relatively modest factor can have a significant effect on running time. It has been suggested that instead of using all samples, or using a random selection of samples as proposed in [21], one may use an efficient representation of training samples by means of dictionary learning methods [24],[23] or use subsets of training samples which efficiently represent each class [8].

A dictionary can be considered as a linear basis where data

has a different (generally sparse) representation. As an example, discrete Fourier transform (DFT) and various types of Wavelet methods can be viewed as unsupervised dictionaries. Recently, supervised dictionary learning has drown interest in the context of machine learning and pattern recognition. In these methods, original training samples are utilized to build a dictionary. Many of dictionary learning approaches can only represents data efficiently but are not designed for classification tasks [1]. On the other hand, dictionaries with classification power can be effectively designed and used in the SRC framework to increase the time and memory efficiency of this approach. In this work, the performance of the SRC algorithm is studied in combination with three recently proposed supervised dictionary learning methods, Metaface learning [24], Fisher Discriminative Dictionary Learning (FDDL)[23] and Sparse Modeling Representative Selection (SMRS)[8]. In this paper the experiments are conducted on a face recognition framework with different datasets and the results are compared with the original SRC method where a random selection of training data are used directly as dictionary elements.
The rest of this paper is organized as follows:. In section 2 the sparse representation classification algorithm is briefly reviewed , In section 3, general dictionary learning method is described and three methods of dictionary learning (Metaface learning, Fisher Discriminative Dictionary Learning (FDDL) and Sparse Modeling Representative Selection (SMRS)) are described. The results of the experiments and the details of the datasets are reported in the section 4. This section also elaborates on how different dictionary learning methods affect SRC accuracy.

## 2. SPARSE REPRESENTATION BASED CLASSIFICATION

The system of linear equations $\mathbf{t} = D\mathbf{s}$, where $D \in \mathbb{R}^{m \times n}$ is an $m$ by $n$ matrix, is called an under-determined system of equations if $m < n$. In this system, the measurement vector $\mathbf{t}$, is a column vector with $m$ entries, and $\mathbf{t}$ which is a column vector with $n$ entries is the vector to be recovered. The solution $\widehat{\mathbf{t}}$ to this equation is not unique and the sparsest solution to this equation can be obtained by solving the following optimization problem:

$$\widehat{\mathbf{s}}_0 = \text{argmin} \left\|\mathbf{s}\right\|_0 \text{ subject to } D\mathbf{s} = \mathbf{t}\,, \tag{1}$$

where $\|.\|_0$ denotes the $ell^0$-norm, which counts the number of non-zero elements in vector $\mathbf{s}$. Finding the solution for $\mathbf{s}$, using (1) is an NP-hard problem because all the subsets of the entries for $\mathbf{s}$ should be considered [2]. Based on the theory of compressive sensing, if the solution $\widehat{\mathbf{s}}_0$ is sparse while satisfying certain constraints [6], the solution of the optimization problem (1) is equal to the solution of the following $\ell^1$-norm minimization problem:

$$\widehat{\mathbf{s}}_1 = \text{argmin} \left\|\mathbf{s}\right\|_1 \text{ subject to } D\mathbf{s} = \mathbf{s}\,. \tag{2}$$

In fact, vector $\widehat{\mathbf{s}}_1$ should not necessarily be sparse to be recovered by (2). It may be sparse in some domain (other than $\widehat{\mathbf{s}}_1$'s original domain) [3]. For instance, vector $\widehat{\mathbf{s}}_1$, could be a general non-sparse signal which has a sparse representation in frequency or Wavelet domain. In practice, due to the existence of noise in measurements, the solution to $\mathbf{t} = D\mathbf{s}$ is not exact. In other words, the system of linear equations, $\mathbf{t} = D\mathbf{s}$ should be modified as $\mathbf{t} = D\mathbf{s} + \mathbf{n}$, where $\mathbf{n}$ is an $n$ dimensional noise vector. In this case, the optimization problem (2) may be reformulated as follows:

$$\widehat{\mathbf{s}}_1 = \text{argmin} \left\|\mathbf{s}\right\|_1 \text{ subject to } \left\|D\mathbf{s} - \mathbf{t}\right\|_2 \le \epsilon\,, \tag{3}$$

where, $\epsilon > \|\mathbf{n}\|_2$ , i.e. $\epsilon$ is larger than the energy of the noise. The idea of classification in the context of sparse representation is to set up a system of linear equations, $\mathbf{t} = V\mathbf{s}$, where $V \in \mathbb{R}^{m \times n}$ represents the original training data which totally contains $n$ samples. Each column of the matrix $V$ is an $m$ dimensional training sample vector $\mathbf{v}$ . Vector $\mathbf{t}$ is an $m$ dimensional test sample which is unknown and $\mathbf{s}$ is the coefficient vector representing the test sample as a linear combination of the training samples. Training samples for all $c$ classes can be considered as $V = [V_1, V_2, \ldots, V_c]$, where each $V_i \in \mathbb{R}^{m \times n_i}$ is a sub-matrix contains $n_i$ training samples which spans a subspace for class $i$, i.e. $V_i = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots, \mathbf{v}_{i,n_i}]$. In an ideal situation, a test sample of this class, $\mathbf{t}_i \in \mathbb{R}^m$ can be represented as a linear combination of the training samples as follows:

$$\mathbf{t}_i = s_{i1}\mathbf{v}_{i,1} + s_{i2}\mathbf{v}_{i,2} + \cdots + s_{in_i}\mathbf{v}_{i,n_i} = V_i\mathbf{s}_i\,, \tag{4}$$

where $\mathbf{s}_i \in \mathbb{R}^{n_i}$ is the coefficient vector for class $i$. Considering all $n = n_1 + n_2 + \cdots + n_c$ training samples from all $c$ classes, the entire dictionary could be represented by the matrix $V \in \mathbb{R}^{m \times n}$ where

$$\begin{aligned} V &= [V_1, V_2, \ldots, V_c] \\ &= [\mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \ldots, \mathbf{v}_{1,n_1}, \mathbf{v}_{2,1}, \ldots, \mathbf{v}_{2,n_2}, \ldots, \mathbf{v}_{c,n_c}]\,. \end{aligned} \tag{5}$$

The linear system shown in (4) can be represented in matrix form as

$$\mathbf{t} = V\mathbf{s}\,, \tag{6}$$

where $\mathbf{s} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_c]^T$. Ideally, the solution to equation (6) for a test sample from class $i$ i.e. $\mathbf{t}_i$ will be $\mathbf{z}_i$ which is a sparse vector whose entries are mostly zero except for the ones corresponding to $i^{th}$ class:

$$\mathbf{z}_i = [0, 0, \ldots, 0, s_{i1}, \ldots, s_{in_i}, 0, 0, \ldots, 0]^T\,. \tag{7}$$

If $m < n$, i.e. the dimension of the data is smaller than the total number of training samples, equation (6) represents an under-determined system. Assuming that the training samples from one class represent a subspace with a lower dimension than $m$, it is possible to recover $\mathbf{s}$ by solving $\ell^1$-norm minimization problem (2). Having the recovered vector, $\mathbf{s}$, it is possible to find the class of the given test sample. This classification method is called Sparse Representation-based Classification (SRC) [21]. In real applications, due to the existence of measurement noise, both test samples and training samples (dictionary atoms) are noisy and it may not be possible to represent a test sample only as a linear combination of training samples which belong to the same class. To address this problem, the constraint on the $\ell^1$-norm minimization problem is changed to consider the noise effect, as described earlier in equation (3). Another consideration in SRC algorithm is the dimensionality of the data. if data dimension is large, then it is necessary to have a large number of training samples for equation (6) to be under-determined. For example, in a face recognition framework, if training data consists of face images with the resolution of $100 \times 100$ , then the training matrix $V$ will have $10^4$ rows which implies that the number of training samples must be greater than $10^4$ to have an under-determined system of linear equations. In order to lower the number of rows in the training matrix, it is possible to reduce the dimensionality of the original data, $m$, into $d \ll m$ using a proper feature extraction algorithm. Many of feature extraction algorithms consist of linear transformations which may be represented as a matrix multiplication. In this case, the equation (6) could be rewritten as follows:

$$\widehat{\mathbf{t}} = R\mathbf{t} = RD\mathbf{s} = \widehat{V}\mathbf{s}\,, \tag{8}$$

where $\widehat{\mathbf{t}} \in \mathbb{R}^d$ represents the extracted feature vector of the original unknown test sample $\mathbf{t}$, and $R \in \mathbb{R}^{d\times m}$ with $d \ll m$, is a feature extraction matrix. $\widehat{V} \in \mathbb{R}^{d\times n}$ is the training matrix with reduced dimensionality. Using $\widehat{V}$ as the training matrix, equation (3) can be reformulated as:

$$\widehat{\mathbf{s}}_1 = \text{argmin}\, \|\mathbf{s}\|_1 \text{ subject to } \left\|\widehat{V}\mathbf{s} - \widehat{\mathbf{t}}\right\|_2 \leq \epsilon\,. \tag{9}$$

Different dimensionality reduction matrices such as random projection and down-sampling can be utilized as $R$ in equation (9).

# 3. DICTIONARY LEARNING METHODS

Dictionary learning (DL) is a method for representing training data to make it more efficient for coding and further processing and it is widely used in different applications such as computer vision and signal processing. In most of these applications, dictionaries are designed such that the training set could be represented as sparse as possible. An effective approach to build a dictionary is to use the training samples in dictionary learning algorithm [17]. For training samples $V$, one could end up with a dictionary $D$ with solving the following optimization problem:

$$\begin{aligned}
&\mathop{\textbf{argmin}}_{D,\mathbf{x}_i} \sum_{i=1}^{n} \|\mathbf{v}_i - D\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_1\\
=&\mathop{\textbf{argmin}}_{D,X} \sum_{i=1}^{n} \|V - DX\|_F^2 + \lambda\|X\|_1\\
&s.t.\ \|\mathbf{d}_i\|_2^2 \leq 1, \forall i = 1,2,\ldots,n,
\end{aligned} \tag{10}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ is the coefficient matrix which can be considered as the representation of training samples $V$ over the dictionary $D$. There are several proposed dictionary learning methods to create dictionaries which are also effective for classification tasks [16],[25].

as described in section 2 an important part of the SRC algorithm is the training matrix $V$ which formed simply by adding the vectorized version of the training samples. Despite of the interesting classification results in [21], SRC would not be computationally efficient if the number of training samples is large. Building a dictionary by manual selection from the training samples as suggested in [21] is also non-optimal, because one might select training samples which are not good representatives for the training data. Considering that many of training samples from one class might be redundant and contain similar information, it is possible to find a more efficient matrix with less elements. If the training matrix $V$ in equation (6) is replaced by a smaller size dictionary $D$ which is as representative and discriminative of the training data, classification performance will be increased while maintaining the accuracy. Different dictionary learning algorithms are used along with SRC. There are two main category of methods presented in this area. Some of these methods do not create a sparse coefficient vector to be used directly for classification, i.e. dictionaries built by this methods are not convenient for classification. Since the recovered vector in SRC algorithm needs some further processing to detect the unknown sample class [16],[25]. In contrast, a number of methods are proposed which try to incorporate the discriminative power of a dictionary into the solution of the optimization function [24],[23],[8]. In this paper, we study the methods from second group and evaluate the performance of these algorithms compared to each other.

## 3.1 Metaface Dictionary Learning

Yang et al. [24] introduced a dictionary learning method based on metagenes in gene expression data analysis, and used it along with SRC algorithm in a face recognition framework. In this approach, dictionary $D$ is considered to be a collection of sub-dictionaries representing each class separately, i.e. $D = [D_1, D_2, \ldots, D_c]$, where each $D_i$ is learned for each class $i$ separately by using training samples from that class, $V_i$. Atoms in the sub-dictionary $D_i$ are denoted by $\mathbf{d}_{i,j}$ and $D_i = [\mathbf{d}_{i,1}, \mathbf{d}_{i,2}, \ldots, \mathbf{d}_{i,p}]$, where $p \leq n$. Each atom in the dictionary required to be a unit vector, i.e. $\mathbf{d}_{i,j}^T\mathbf{d}_{i,j} = 1, \forall i, j$. Metaface dictionary will be determined by solving the following optimization problem:

$$\mathop{\textbf{argmin}}_{D_i,X_i} \|V_i - D_iX_i\|_F^2 + \lambda\|X_i\|_1\ s.t.\ \mathbf{d}_{i,j}^T\mathbf{d}_{i,j} = 1, \forall i, j, \tag{11}$$

where $X_i$ is the representation of the training data $V_i$ over the sub-dictionary $D_i$. $\lambda$ is the regularization parameter and $\|.\|_F$ indicates the Frobenius norm of matrices. Equation (11) is a multi-variable optimization problem and could be solved by alternatively optimizing $D_i$ and $X_i$. The overall optimization steps is as follows (details for each step are described in [24]):

***Step 1:*** Each column of $D_i$ is initialized as a random vector with unit $l_2$ norm.
***Step 2:*** Fix $D_i$ and solve for $X_i$.
***Step 3:*** Fix $X_i$ and update $D_i$.
***Step 4:*** If maximum number of iterations or an acceptable error rate are not met, return to step 2

The above algorithm should be conducted for each individual class $i$ ($i = 1, 2, \ldots, c$) to form the final dictionary $D = [D_1, D_2, \ldots, D_c]$. This dictionary could be used in SRC framework to perform the classification task.

## 3.2 Fisher Discrimination Dictionary Learning

Another study on using dictionary learning along with SRC framework is done in [23]. The so-called Fisher Discrimination Dictionary Learning (FDDL) tries to learn a dictionary which contains atoms with class labels and in the meantime, utilizes Fisher discrimination criterion to make the dictionary more discriminative. To satisfy this property, FDDL uses the following optimization problem:

$$\mathop{\textbf{argmin}}_{D,X} \left\{r(V,D,X) + \lambda_1\|X\|_1 + \lambda_2 f(X)\right\}, \tag{12}$$

where $V = [V_1, V_2, \ldots, V_c]$ and $X = [X_1, X_2, \ldots, X_c]$ are the training and coefficient matrices for all classes $(1, 2, \ldots, c)$ respectively, and $D$ is the final dictionary as described in section 3.1. $\lambda_1$ and $\lambda_2$ are regularization parameters, $r(V, D, X)$ is the discriminative fidelity term, and $f(x)$ is the Fisher discrimination constraint term. Discriminative fidelity term imposes the following 3 constraints on the dictionary and coefficient matrix for class $i$ ($i = 1, 2, \ldots, c$):

***C1:*** The dictionary $D$, should be a good representative of the training samples from class $i$, i.e. $V_i$ with coefficient matrix $X_i$.
***C2:*** Sub-dictionary $D_i$ along with coefficient entries associated with class $i$ ($X_i$), should be a good representative for $V_i$.
***C3:*** Coefficient entries associated with class $j \neq i$ ($X_i^j$), should be close to zero to make $D_jX_i$ small.

Mathematical interpretation of the above constraints results in the following formula for discriminative fidelity term:

$$r(A_i, D, X_i) =$$

$$\|A_i - DX_i\|_F^2 + \left\|A_i - D_iX_i^i\right\|_F^2 + \sum_{\substack{j=1 \\ j\neq i}}^{c} \left\|D_jX_i^j\right\|_F^2 . \quad (13)$$

Fisher discriminative criterion increases dictionary discriminative power by minimizing the within-class scatter and maximizing the between-class scatter of $X$. Based on this criteria, the discriminative term, $f(X)$ in equation (12) for class $i$ can be formulated as equation (14) (Details are provided in [23]).

$$f_i(X_i) = \|X_i - M_i\|_F^2 - \sum_{k=1}^{c} \|M_k - M\|_F^2 + \|X_i\|_F^2 , \quad (14)$$

where $M_k$ and $M$ are the mean vector matrices with $n_k$ columns whose columns are mean vectors of $X_i$ and $X$ respectively. The optimization functions (13) and (14) are convex [23], so the optimization function (12) can be solved by a similar approach of 3.1, i.e. alternatively optimization of $D$ and $X$.

***Step 1:*** Each column of $D_i$ is initialized as a random vector with unit $l_2$ norm.
***Step 2:*** Fix $D_i$ and solve for $X_i$ for all classes ($i = 1, 2, \ldots, c$).
***Step 3:*** Fix $X_i$ and update $D_i$ for all sub-dictionaries ($i = 1, 2, \ldots, c$).
***Step 4:*** If maximum number of iterations or an acceptable error rate are not met, return to step 2

The solution to (12) will result in a discriminative structured dictionary whose atoms are classified and labelled for their associated class. This dictionary is used as the training matrix in SRC algorithm.

## 3.3 Sparse Modeling Representative Selection

While Metaface and FDDL methods try to build a new dictionary by processing and modifying the training data, the Sparse Modeling Representative Selection (SMRS) proposes to form the dictionary by selecting its atoms from the original training samples. Representatives for each class $i$ of training data, are selected using the following optimization problem:

$$\underset{X_i}{\mathbf{argmin}} \lambda \|X_i\|_{1,q} + \frac{1}{2} \|V_i - V_iX_i\|_F^2 , \ s.t. \ 1^\top X_i = 1^\top , \quad (15)$$

where $\|X_i\|_{1,q} = \sum_{j=1}^{n_i} \|X_i(j,:)\|_q$, where $\|X_i\|_{1,q}$ is the sum of $\ell^q$-norms of the rows of $C_i$, and $q$ is selected to be greater than one to make the optimization problem convex [8]. Scalar $\lambda$ is the regularization parameter which affects the number of selected representatives and the constraint term $1^\top X_i = 1^\top$, enforces the solution to be translation invariant. The optimization problem (15) can be solved using Alternating Direction Method of Multipliers (ADMM) method from [9]. The selected representatives, which are directly used as dictionary atoms, are the columns of training samples $V_i$ which have non-zero corresponding row in the coefficient matrix $X_i$.

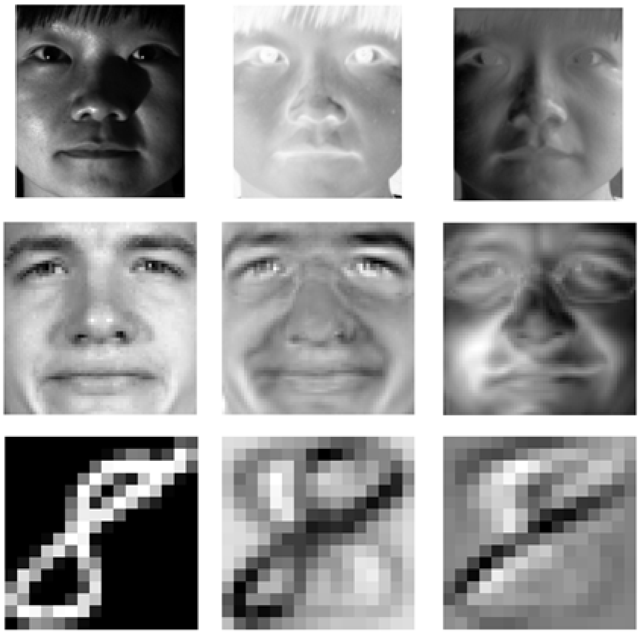

**Figure 1: SMRS(left), Metaface(middle) and FDDL (bottom) representatives selected on subjects 3, 4514 and 8 from Yale B (top), FRGC (middle), and Cedar Buffalo (bottom) datasets.**

The above scheme also indicates which samples are more informative to represent class $i$ by directly comparing $\ell^q$-norms of the selected samples. Moreover, it is possible to detect outliers by checking the coefficient matrix $X$. The rows of $X$ which corresponds to outliers should have a few non-zero entries. Based on this fact, a measure called row-sparsity index ($rsi$) is introduced in [8] which is used for removing outlier representatives. Some discussions are also made in [8] on how to efficiently update the representatives for each class when new training samples are in introduced.
Figure 1 shows a column of SMRS, Metaface and FDDL dictionaries for three different datasets which are introduced in section 4. As can be seen, SMRS dictionary (left column) selected some of the original data as representatives. On the other hand, Metaface and FDDL methods, learn their own representatives which are different from original images.

# 4. EXPERIMENTS

In this section, the classification performance of the SRC method is evaluated using different dictionaries and datasets. In order to compare the results of the three dictionary methods, described in the previous section, the number of atoms in the dictionaries should be equal. The Metaface and the

**Table 1: Recognition rate (using down-sampling(DS) and random projection(RP)) and learning time of using different dictionary learning methods on Extended Yale B face dataset**

|  | Accuracy<br>% (DS) | Accuracy<br>% (RP) | Learn Time<br>(sec) | Test Time<br>(sec) |
|---|---|---|---|---|
| Random | 85.79 | 92.49 | N/A | 0.051 |
| SMRS | 91.53 | 93.25 | 20 | 0.050 |
| Metaface | 86.60 | 88.17 | 1300 | 0.051 |
| FDDL | 92.52 | 94.04 | 19400 | 0.049 |
| All Data | 97.70 | 98.37 | N/A | 0.390 |

FDDL methods allow the user to directly control the number of dictionary atoms. However, in the SMRS method the number of dictionary atoms are affected by the nature of the training data and the parameter $\lambda$ in the equation (15). Hence, we first run the SMRS method with a given parameter $\lambda$ on all classes. In practice, a second parameter $\alpha > 0$ is selected to compute $\lambda$ via $\lambda = \lambda_0/\alpha$ where $\lambda_0$ depends on the nature of the data [8]. ADMM method is employed to solve the optimization problem (15), as described in section 3.3.

Once the SMRS method selects the number of representatives for each class, then the same number of atoms is used in the second and third experiments with the other two dictionary learning methods, i.e. Metaface and the FDDL. The $\ell^1$-regularized least square problem (11) is alternatively optimized for dictionary and coefficients by the interior-point method presented in [14] and optimization problem (12) is solved alternatively using Metaface approach [24] and iterative projection method presented in [19]. In order to compare the results of the three dictionary learning methods with the original SRC method, the forth experiment was performed where the dictionary atoms were randomly selected directly from the training set and used for the classification task. This experiment, i.e. the random selection of the dictionary atoms from the original training data, was independently repeated for 10 times and an average recognition rate was calculated. Finally, in the fifth experiment, all the training images were used directly to form the largest possible dictionary for classification. The results of this last experiment was compared to the first four experiments.

different popular datasets, Cedar Buffalo digits, FRGC, and Yale B face datasets, were used to compare the performance of the aforementioned dictionary learning methods. The results are reported in the following sections.

## 4.1 Yale B Face Dataset

The first face dataset used for the experiments were selected from the Extended Yale B face dataset [15], [10]. This dataset contains a total of 2414 face images from 38 subjects which are cropped and normalized into 192 × 168 frontal face images. Images are captured under various controlled lighting conditions in the laboratory. Half of these images (1207) were selected randomly for training and the remaining were used as test samples. In the first experiment, SMRS algorithm with a fixed predetermined $\lambda$ for all classes was applied on training images to select the representatives for the dictionary. This algorithm selected 8,9,10 or 11(with an average number of 9.58) representatives for each class (364 total representatives). Total running time for SMRS algorithm to build the dictionary was around 20 seconds. After this step, original image vectors of length 32256 were down-sampled into 120 dimensional vectors. Recognition rate using these representatives in an SRC framework is 91.53% while the average classification time for an unknown input is 50.87 milliseconds. The same number of representatives for each class were forced for the Metaface dictionary learning method while regularization parameter was selected to be $\lambda = 0.001$. Total running time for this algorithm to build the dictionary was around 1300 seconds which is far larger than the time to learn SMRS dictionary. Recognition rate for SRC using a down-sampled version of the Metaface dictionary (similar approach to the first experiment) was 86.60% which shows a 5% decline in recognition rate comparing to the SMRS dictionary learning method. Since SRC for both methods uses a dictionary of the same size, the average single classification time for both SMRS and Metaface methods are comparable, 50.87 and 51.12 milliseconds correspondingly.

The experiment was repeated using FDDL dictionary learning method with the same number of representatives for each class. The process time for training the FDDL dictionary was 19400 seconds which was much more than the learning time of the other two methods. For the same feature extraction approach (i.e. down-sampling), SRC recognition rate was 92.52% which is the best accuracy among all three methods.

In order to test the effect of feature extraction on classification results, the above experiments were repeated using a random projection feature extraction which projects face images into a 120 dimensional space. SRC was repeated 10 times with different random projection matrices and the average recognition rate using SMRS, Metaface and FDDL dictionary learning methods were 93.25%,88.17% and 94.04%. Table 1 shows the summary of the learning and classification results for the Extended Yale B dataset.

## 4.2 FRGC Face Dataset

The second face dataset used for the experiments consists of 2D front face images from the FRGC dataset [18]. Images in this dataset are different from Yale B dataset because rather than having various lighting conditions, they are captured in different times, poses and situations. This dataset contains 36817 face images from 535 subjects (i.e., 535 classes). Among all classes, 100 classes were randomly selected for the experiments. The original resolution of images was either 1704 × 2272 or 1200 × 1600. All images were converted to gray images, normalized and cropped to 60 by 60 pixels. For each class, 80 and 30 face images were randomly se-

**Table 2: Recognition rate (using down-sampling(DS) and random projection(RP)) and learning time of using different dictionary learning methods on FRGC face dataset**

|  | Accuracy<br>% (DS) | Accuracy<br>% (RP) | Learn Time<br>(sec) | Test Time<br>(sec) |
|---|---|---|---|---|
| Random | 90.23 | 83.10 | N/A | 0.32 |
| SMRS | 94.30 | 85.76 | 38 | 0.32 |
| Metaface | 90.77 | 80.65 | 8200 | 0.31 |
| FDDL | 94.10 | 88.02 | 91000 | 0.34 |
| All Data | 96.80 | 97.70 | N/A | 15.14 |

**Table 3: Recognition rate (using down-sampling(DS) and random projection(RP)) and learning time of using different dictionary learning methods on Cedar Buffalo digit dataset**

| | Accuracy % (DS) | Accuracy % (RP) | Learn Time (sec) | Test Time (sec) |
|---|---|---|---|---|
| Random | 82.49 | 76.20 | N/A | 0.02 |
| SMRS | 88.82 | 80.83 | 85 | 0.02 |
| Metaface | 85.93 | 79.40 | 5377 | 0.02 |
| FDDL | 90.16 | 85.08 | 2298 | 0.02 |
| All Data | 97.50 | 95.58 | N/A | 5.22 |

lected as training and testing sets respectively which results in a total number of 8000 training images and 3000 test images (images in testing set were selected to be different from training images).

SMRS algorithm was employed first to select training representatives and form the first dictionary. The parameter $\lambda$ was selected such that the average number of representatives was 12.5 with the total dictionary learning running time of 38 seconds. The same number of representatives were forced to Metaface and FDDL dictionary learning methods which had far longer learning processes (8200 and 91000 seconds respectively).

Learned dictionaries using the three methods were used along with a down-sampling feature extraction matrix for classification. This dimensionality reduction changes sample vectors length from 3600 to 100. Recognition rates using SMRS, Metaface and FDDL dictionaries were 94.30%, 90.77% and 94.10%, respectively. For this dataset, similar to Yale B dataset, SRC accuracy using SMRS and FDDL dictionaries performed better comparing to when Metaface dictionary was employed. The average recognition rate were 85.76%, 80.65% and 88.02% for SMRS, Metaface and FDDL dictionaries respectively. The results of using different dictionary learning methods for classification of FRGC face dataset are summarized in Table 2.

## 4.3 Cedar Buffalo Digits Dataset

To evaluate different dictionary learning methods in another context, a set of experiments were conducted on handwritten digits dataset from the Cedar Buffalo binary digits dataset (USPS) [13]. This dataset contains 11000 examples of 8 bits, 16 × 16 digit bitmaps (1100 images for each of digits $0, 1, \ldots, 9$). Among these images, Half of the images in each class were selected as training samples and the rest were used to test the classification algorithm.

At the first step, SMRS dictionary learning was employed to build the dictionary matrix. Parameter $\lambda$ was selected to create an average of 24.7 representatives per class and the dictionary learning running time was 85 seconds. Similar to face datasets, Metaface and FDDL dictionary learning methods were applied with the same number of representatives per class as SMRS dictionary. Dictionary learning running time was 5377 and 2298 seconds for Metaface and FDDL learning methods respectively.

Down-sampling and random projection were used to reduce dimensionality of the data from 256 to 64. Recognition rates using down-sampled SMRS, Metaface and FDDL dictionaries were 88.82%, 85.93% and 90.16% respectively while the average recognition rate over 10 runs of random projection dimensionality reduction were reported as 80.83%, 79.40% and 85.08% respectively. An average testing time of 24 milliseconds was reported for each digit test sample in the above classification framework. Table 3 shows recognition rate and learning time of SMRS, Metaface and FDDL dictionary learning methods as well as using all and random selection of training images for the digit recognition problem.

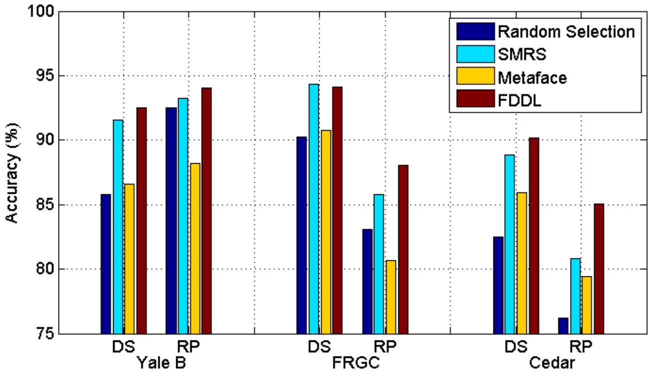

**Figure 2: Recognition accuracy on different datasets using dimensionality reduction methods down-sampling(DS) and random projection(RP).**

Figure 2 shows SRC accuracy using different dictionary learning and dimensionality reduction methods on the three selected datasets. As can be seen in this figure, the SRC has the best recognition rate when the FDDL learning method is used to form the dictionary. Further discussions on the results is presented in section 5.

## 5. DISCUSSION AND CONCLUSION

This paper investigated the influence of the three different dictionary learning methods on the performance of the sparse representation-based classification (SRC). The original SRC algorithm which is proposed in [21], works based on the assumption that a test sample from a particular class can be represented as a linear combination of the training samples from the same class. Therefore, an unknown test sample can be recognized by recovering the sparse coefficient vector which is a representation of the test sample over training images from all classes. $\ell^1$-norm optimization is utilized to solve this reconstruction problem. Using all the training samples makes the optimization process slow and a random selection from the training samples is also inefficient. To address this problem, dictionary learning methods can be utilized to reduce the number of representatives for each class. Dictionary learning methods which are evaluated in this study include Metaface, Fisher Discriminative Dictionary Learning (FDDL) and Sparse Modeling Representative Selection (SMRS). SMRS method selects the best representatives directly from training dataset while the other two methods build their representatives for each class by processing the original training samples while optimizing an objective function. The learned dictionaries from the three aforementioned methods were used to feed the SRC algorithm for classification of two face and one digit datasets. The accuracy and the learning performance of these methods were compared. Two dimensionality reduction algorithms (down-sampling and random projection) were also used in order to make a better comparison of the accuracy of the methods. From the learning point of view, SMRS was the fastest method. While Metaface and FDDL methods needed more than an hour to learn the dictionary, the learning time for SMRS was less than a minute. This difference in the learning phase make SMRS method the best choice for dynamic situations where the dictionary is regularly updated with new samples. While FDDL method introduced the longest learning process on Yale B and FRGC face datasets, Metaface learning time was the largest one in the digit dataset. This fact shows that Metaface learning time highly depends on the number of training images per class. On the other hand, FDDL needs more time to learn when number of classes and the dimensionality of the data is higher.

Investigation of the results for the three different dictionary learning methods show that selecting the FDDL leads to the best recognition rate for the SRC method (except for one case where SMRS had 0.2% better recognition rate than FDDL on FRGC face dataset). The recognition rates of using the SMRS were generally slightly lower than the FDDL. The Metaface dictionary learning method accuracy was specifically less than the other two and even in some cases it was worse than a simple random selection of the training data as the dictionary (Using random projection on Yale B and FRGC datasets). As expected, using all the training samples as the dictionary resulted in the best recognition rates in all the datasets but the classification times for this approach were far larger than the required time for any one of the three dictionary learning methods.

Analysis of the results for the down-sampling and random projection feature extraction methods show that for the FRGC and Cedar datasets, down-sampling features were more effective than random projection using all 4 representative selection methods but this is not the case for Yale B dataset. This difference may be the result of the characteristics of these datasets. Yale B dataset contains face images which are similar in pose and expressions but only captured in different controlled lighting conditions while the other two datasets were not captured within a controlled environment.

As a summary, one can conclude that using learned dictionaries in an SRC framework, leads to faster classification process comparing to when all training images are used to form the SRC linear system of equation. Among the selected dictionary methods, FDDL introduced the highest recognition rate while its learning time was much higher than SMRS. This makes FDDL to be more applicable in offline applications and SMRS to be more suitable in dynamic learning applications.

## 6. REFERENCES

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

[2] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 1998.

[3] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[4] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces versus fisherfaces: Recognition using class specific linear projection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[5] E. Candes. Compressive sampling. *International Congress of Mathematics*, 2006.

[6] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[7] D. Donoho and Y. Tsaig. Fast solution of l1-norm minimization problems when the solution may be sparse. *Department of statistics, Stanford University*, 2006. "available at: http:/dsp.rice.edu/sites/dsp.rice.edu/files/cs/FastL1.pdf".

[8] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. pages 1600–1607, 2012.

[9] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1):17–40, 1976.

[10] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transaction Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[11] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[12] K. Huang and S. Aviyente. Sparse representation for signal classification. *Advances in neural information processing systems*, 19:609, 2007.

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

[14] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(8):1519–1555, 2007.

[15] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

[16] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *preprint - arXiv:0809.3083*, 2008.

[17] J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *preprint - DTIC Document*, 2007.

[18] P. Phillips, P. Flynn, T. Scruggs, and K. Bow. Overview of the face recognition grand challenge. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[19] L. Rosasco, A. Verri, M. Santoro, S. Mosci, and S. Villa. Iterative projection methods for structured sparsity regularization. *MIT Computer Science and Artificial Intelligence Laboratory Technical Report (MIT-CSAIL-TR-2009-050)*, 2009.

[20] M. Turk and A. Pentland. Eigenfaces for recognition. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1991.

[21] J. Wright, A. Yang, A. Ganesh, and S. Sastry. Robust face recognition via sparse representation. *IEEE Transaction on Pattern Analysis and Machine*

*Intelligence*, 31(2):210–227, 2009.

[22] A. Yang, S. Iyengar, S. Sastry, R. Bajcsy, P. Kuryloski, and R. Jafari. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[23] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. pages 543–550, 2011.

[24] M. Yang, L. Zhang, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. pages 1601–1604, 2010.

[25] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. pages 2691–2698, 2010.