スパース表現分類のための典型事例選択手法

AliceChen[†,††]　江田　毅晴[†]　片渕　典史[†]　片岡　良治[†]

† NTT サイバーソリューション研究所, 日本電信電話株式会社
〒 239-0847 神奈川県横須賀市光の丘 1-1
†† School of Computing Science, Simon Fraser University
8888 University Drive, Burnaby BC, V5A 1S6 Canada
E-mail: †{alice.chen,eda.takeharu,katafuchi.norifumi,kataoka.ryoji}@lab.ntt.co.jp

あらまし　画像中の顔認識はパターン認識分野において長く研究されてきたが，既存手法の多くはロバスト性が不足していたり計算量が高いという課題があった．近年，圧縮センシング理論に基づいたスパース表現分類法 (SRC: Sparse Representation-based Classification) [12] と呼ばれる新しい手法が, ノイズを含む画像の顔認識に有効との研究例が報告されている．SRC では，未知データは訓練データ中の全てのアイテムの線形結合で表現されると考える．その表現がスパース性を満たすという仮定のもと，圧縮センシング理論に基づき未知データのスパース信号を効率的に復元し，復元信号が所属するクラスを求めることによって，顔認識を行う．しかしながら，SRC では訓練データを用いて作成するセンシング行列の大きさに比例して分類に時間がかかるため，現実的な問題に適用するのは困難であった．そこで本研究では，訓練データ中の典型例を選択するための基準を提案する．これにより，正解分類により貢献するアイテム集合のみでセンシング行列を構築することで，認識精度の低下少なく時間・空間的に効率的な顔認識が可能となる．
キーワード　スパース表現分類 (SRC)，スパース表現，クラス分類，パターン認識，顔認識，効率

A Representative Sample–Selection Approach for SRC

Alice CHEN[†,††], Takeharu EDA[†], Norifumi KATAFUCHI[†], and Ryoji KATAOKA[†]

† NTT Cyber Solutions Laboratories, NTT Corporation
1–1 Hikarinooka, Yokosuka, Kanagawa, 239–0847 Japan
†† School of Computing Science, Simon Fraser University
8888 University Drive, Burnaby BC, V5A 1S6 Canada
E-mail: †{alice.chen,eda.takeharu,katafuchi.norifumi,kataoka.ryoji}@lab.ntt.co.jp

Abstract　Face recognition has been a popular topic in pattern recognition due to the different applications it can be applied to. Traditional methods for face recognition have suffered from the lack of robustness as well as high computational cost. Recently, SRC (Sparse Representation-based Classification) [12], a new approach based on concept from compressive sensing theory, has been shown to be more efficient and effective in dealing with noisy images. In SRC, each sample is represented as a linear combination of all the items in the training set; such representation is sparse, therefore can be recovered using reasonably efficient methods. However, speed performance generally degrades as the size of training set increases. In this paper, we propose a technique that can automatically select a subset of representative items from the training set based on a set of validated samples, and is able to improve the time and space efficiency of the recognition task without losing accuracy significantly.
Key words　SRC, classification, sparse representation, pattern recognition, face recognition, efficiency

1. Introduction and Motivation

A typical face recognition (FR) system is composed of three main tasks, namely 1) face detection, 2) feature extraction and 3) classification. Although face detection and segmentation are difficult problems, they have been well studied and numerous tools can be used to perform the tasks. Traditionally, the dominant approach for the extraction process is via Eigenfaces [11] based on the Principle Component Analysis (PCA) technique; and once the feature space

is constructed, an input sample can be identified using a classifier such as NN (Nearest Neighbors) or SVM (Support Vector Machine). Newly developed methods such as SIFT [8]/SURF [1] feature extraction have been shown to be more robust against varying poses and illumination. However, all these methods essentially rely on extracting specific information from the individual faces. As a result, it is critical to have the correct features and enough of them. This dependency greatly restricts the scalability and practicability of a FR system. For instance, a typical SURF feature point is a vector of length 128 and each face is composed of 200 to 2000 such features [8]. The computation becomes increasingly inefficient as the size of the database increases. Furthermore, high resolution is usually needed to obtain relevant information from a face image. Besides the problems stated, due to the countless possibilities of varying viewing angles and expressions of human faces, accurate recognition is usually limited to faces with frontal views with little or no occlusion using the current FR technology.

These challenges have motivated the development of a new technique for face recognition termed SRC (Sparse Representation-based Classification) [12], which has several benefits over the traditional approach. In contrast to traditional feature extraction methods, in SRC, the role of feature is no longer deterministic as long as the feature space is large enough with sufficient randomness to construct the sparse signals [13]. Due to the sparsity property, it has also been shown that SRC is able to handle noisy images or faces with varying expressions and poses significantly better than other methods [4] [10]. In addition, even data with relatively low resolution can perform reasonably well [12], which makes SRC a more practical model for robust face recognition.

Despite the promising advantages of SRC, the speed performance may still not be satisfactory for many modern face recognition systems, mainly due to the large amount of training data required. In this paper, we will show that by selecting a subset of representative items in the training set, we can greatly improve both time and space efficiency without losing the original accuracy significantly.

2. Background and Related Work

2.1 Compressive Sensing

In signal processing, compressive sensing is the process of reconstructing a signal with the prior knowledge that it is sparse. Mathematically, given an unknown S-sparse signal $x \in \mathbb{R}^n$ for which $\|x\|_0 < S$, compressive sensing tries to recover the signal from its measured data $y \in \mathbb{R}^m$ by finding the most sparse solution to the underdetermined linear system $y = Ax$, where $A = [a_1, a_2..., a_n] \in \mathbb{R}^{m \times n}$ is the sensing matrix with $m < n$. This sparse representation can be recovered by solving the following l1-minimization problem:

$$min\|x\|_1 \text{ subject to } Ax = y$$

In reality, data are usually noisy, so it may not be possible to express a recovered signal exactly, so the problem can be modified to be:

$$min\|x\|_1 \text{ subject to } \|Ax - y\|_2 < \epsilon$$

where ϵ gives a bound to the amount of noise we allow in the data [3] [5] [12].

2.2 SRC

In the case of SRC for face recognition, each image is represented by a column vector composed of the pixel values from the image. The number of pixels to extract from the individual image m corresponds to the dimension of the vector. As described in [12], the sensing matrix A is then constructed by concatenating n of these column vectors, each of which belongs to one of the k classes in the training set. Thus, $A = [A_1, A_2, \ldots, A_k] \in \mathbb{R}^{m \times n}$ for class $1 \ldots k$, with $A_i = [a_i^1, a_i^2, \ldots, a_i^{n_i}]$, where n_i is the number of items in class i.

Geometrically, each image can be thought of as a point in a high dimensional space, and all the training and testing samples lie in this feature space. Then, given a test sample $y \in \mathbb{R}^m$ from one of the classes in the training set, we first use l1-minimization to compute its sparse representation $x \in \mathbb{R}^n$, which is a vector of coefficients. Ideally, the nonzero entries in x will all be the coefficients corresponding to the items from a single class; in that case, we can identify the class y belongs to directly. However, with the presence of noise, such method may not always lead to the correct class assignment.

Alternatively, we can classify y based on how well the coefficients associated with the training samples in each class can reconstruct y. More specifically, for each class i in the training set, we generate a new vector from x defined as $\delta_i(x)$, which consists of all zeros except the entries associated with class i. Then, we assign y the class that minimizes the residual between y and y', where $y' = A\delta_i(x)$. The details of the SRC algorithm is outlined in [12], and 図 1 illustrates the steps in SRC.

In contrast to Nearest Neighbor, which simply classifies a test sample based on a single nearest training sample, SRC considers all training data in each class, which gives it more flexibility to work with a wider range of data. It can effectively avoid problem of under-fitting by utilizing images in different classes to extrapolate the image of interest, instead of the nearest neighbor; and it does so by using the smallest possible set of nonzero coefficients, thereby avoiding the problem of over-fitting [13].

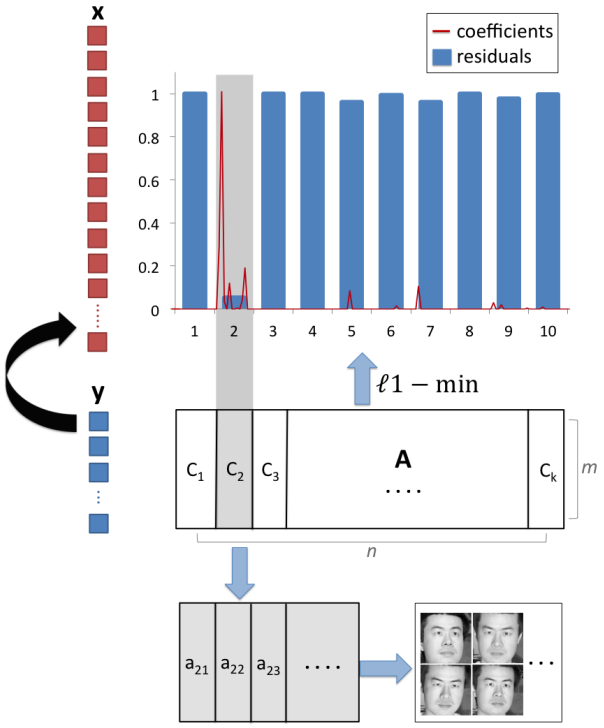


図 1 The input of SRC includes the sensing matrix A and a column vector y representing a test image. Then, l1-minimization gives the coefficient vector x, which is a sparse vector with most of the nonzero entries concentrated in one class. SRC then identifies the class y belongs to based on the minimum residual calculated.

3. Proposed Method

Despite the numerous advantages of SRC, certain conditions limit its applicability in practical applications. Since the sensing matrix A is assumed to be underdetermined, the number of columns n, must be greater than the number of rows m. In fact, it is theoretically desirable to have $m << n$ [5], implying that the size of the training set needs to be large relative to the length of each signal. Moreover, if the variability is small in the training set, then the linear combination of a test sample using the training images is more likely to result in nonzero entries in the recovered vector that are associated with incorrect classes.

Besides the structural restriction on A, performance of the l1-minimization problem is another major issue. In general, the time complexity of solving l1-minimization is $O(n^3)$ [13]. More efficient methods exist, such as Gradient Projection for Sparse Reconstruction (GPSR) [6] or Homotopy [9] algorithms, which can recover solutions that are S-sparse in $O(n+S^3)$ [13] [14], linear in the size of the training set. Nevertheless, it is desirable to keep the complexity as low as possible, especially for real-life face recognition systems that require high speed performance and scalability. In addition, due to the high noise level in many real-life situations, the signals recovered using the existing algorithms often contain many non-zero small coefficients that causes more costly computation.

Algorithm 1 RSS Algorithm

Input: The coefficient matrix G, generated by concatenating each coefficient vector from the validated sample set.
$G = [x_1^1, x_1^2, \ldots, x_1^{t_1}, \ldots, x_k^{t_k}] \in \mathbb{R}^{n \times t}$, where $x_i^j \in \mathbb{R}^n$ is the coefficient vector of the j^{th} validated sample in class i for $i = 1 \ldots k$, and t is the number of validated samples;
p = number of items to select in each class for the reduced matrix.

1: Normalize each column of G to generate the contribution matrix B, where

$$B_c = \frac{\delta_i(x_i^j)}{\|x_i^j\|_1}, \text{ for } c = 1 \ldots t$$

2: Since each row of B corresponds to all the normalized coefficients associated with a single training item, we assign score to each training item a in A by taking the summation of each row in B:

$$Score(a_r) = \|(B^T)_r\|_1 \text{ , for } r = 1 \ldots n$$

3: For each class i, select the top p representative items in A_i based on the calculated score.

$$\arg\max_{a \in C_i} Score(a)$$

Output: A reduced matrix $A' \in \mathbb{R}^{m \times n'}$, where $n' = p \times k$.

By looking at the output of l1-minimization for a set of test samples, we can notice that in general, for a test sample belonging to class i, far fewer samples are needed to represent it than the number of available training items in the class. Also, low resolution images are enough to represent an image to ensure that the vector dimension m does not exceed the training set size n. Also, as stated in the original CS paper [3], the smaller the coherence, the fewer samples are needed to represent a signal. Thus, despite the initial condition on the dimension of the sensing matrix $m << n$, we can relax this constraint with enough variability within the same class to cover the range of possible input, while maintaining high incoherence between different classes.

These observations motivate the idea to utilize the information we get from l1-minimization, which consists of a set of sparse coefficient vectors, for a set of pre-validated samples to reduce the number of columns in the sensing matrix A, thereby improving the speed performance for new input. In other words, once we have enough validated data that roughly cover the range of possible variations of an individual, it is sufficient to select those representative samples that are more likely to contribute to the recovered signal for each class, resulting in a reduced projection from the original A. Algorithm 1 outlines the proposed method, which we call the RSS (Representative Sample-Selection) algorithm.

Basically, we first normalize each coefficient to determine how much each training item contributes to the representation of the original image belonging to class i, relative to all the other training data. Then for each training item in class i, we take the summation of all the coefficients associated with it from all the validated sample data belonging to class i and assign the result as the score. Finally, we can rank the scores and select the top p items from the training set as our representative samples for class i. 図 2 is a visualization of matrix B, which is composed of the concatenation of all the $\delta_i(x)$ column vectors for each validated sample in each class i.

$$B = \begin{pmatrix} C_1 & & & \\ & C_2 & 0 & \\ & & \ddots & \\ & 0 & & C_k \end{pmatrix}$$

図 2 Matrix visualization: each C_i is a block matrix containing the nonzero coefficients associated with class i, and each row corresponds to all the coefficients associated with single training item in A_i.

4. Experiments

4.1 Experimental Setup

We used a subset of the Yale Face Database B [7] for our experiments, which is composed of 10 individuals, or classes, each containing various poses and viewing conditions, with a total of 4500 gray-scale images. We randomly split this dataset into three sets, each of which contains 1500 images. Two of the sets are used for training and generating the reduced training set using RSS, and the remaining set is used for testing the performance of our method. We cross-evaluated the different sets and used the average as the final results. The results we are interested in are the accuracy or recognition rate, and the average processing time of the test sample for running l1-minimization.We also compared our method against random selection method, where a sensing matrix of the same dimension is generated by randomly selecting items from the original training set.

OpenCV library [2] is used to detect and segment faces from the dataset. Each face image is then preprocessed via histogram equalization and down-sampled to a resolution of 10×10. As a result, each image is represented as a single column vector by stacking the 100 8-bit pixel values. Since each set has a length of 1500, the dimension of the sensing matrix A is 100×1500.

The tool used for solving the l1-minimization problem is GPSR_BB(注1), which has been shown to be reasonably efficient and scalable [6] [14]. The output of l1- minimization for those samples that are correctly classified, which we call the validated set, is used as the input of RSS.

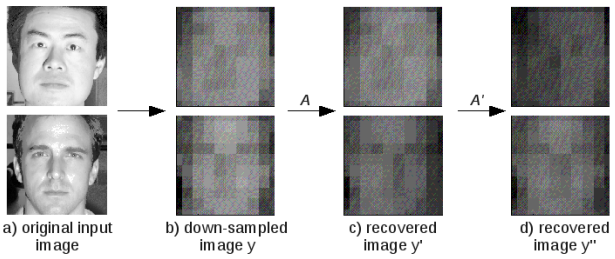


図 3 Two examples of the transformations from a) the original segmented face image to b) the down-sampled 10×10 image y, c) the recovered image y' via l1-minimization with the sensing matrix $A \in \mathbb{R}^{100\times1500}$, where $y' = A\delta_i(x)$, and d) the recovered image y'' when n is reduced to 150, where $A' \in \mathbb{R}^{100\times150}$.

We tested the method using different values of p, where p is the number of representative samples in each class. For $p = [15, 25, 40, 50, 75]$, the dimension of the sensing matrix becomes $100 \times 150, 100 \times 250$, and so on. 図 3 shows some examples of the face images through the down-sampling and recovering process via l1-minimization for the original sensing matrix A and the reduced matrix when $p = 15$.

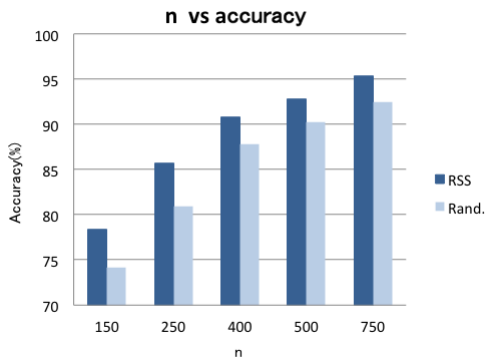
4.2 Results and Comparison

n	Accuracy (%)		Time(s)	
1500	96.87		4.1965	
	RSS	Rand.	RSS	Rand.
750	95.42	92.49	2.0897	2.2207
500	92.81	90.26	1.5947	1.6764
400	90.81	87.80	1.3583	1.4241
250	85.73	80.89	1.0161	1.0597
150	78.37	74.16	0.7702	0.8146

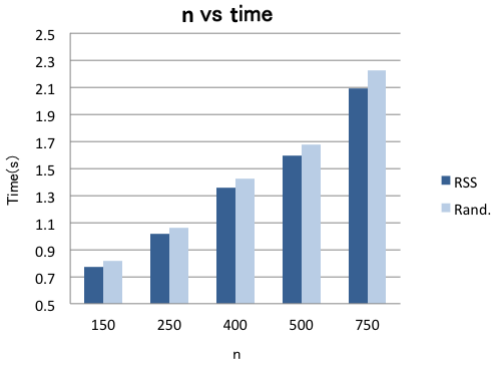
表 1 This table shows the results of comparing our Representative Sample-Selection (RSS) method versus the Random-Selection Approach

Looking at the results in 表 1 and the graphs in 図 4, we can see that the time complexity is linear in respect to the number of columns in A. When we reduce n by half, the accuracy drop merely 1.45%, while the processing time is twice as fast. Even when n has been reduced 10-fold, we still retain 80% of the original accuracy. In other words, only 15 training data is needed in each class to recognize up to 80% of faces out of the 1500 faces in the test set. These results illustrate the power of the SRC in comparison to traditional approaches for FR.

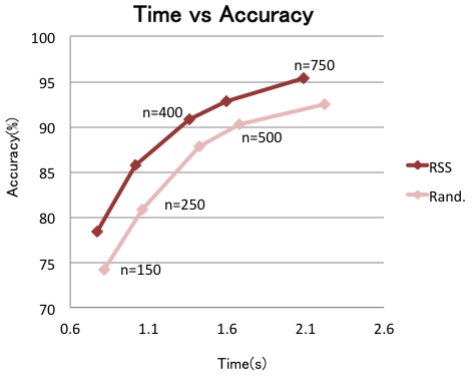
(注1)：The source code for the GPSR_BB algorithm can be found at http://www.lx.it.pt/%7Emtf/GPSR/



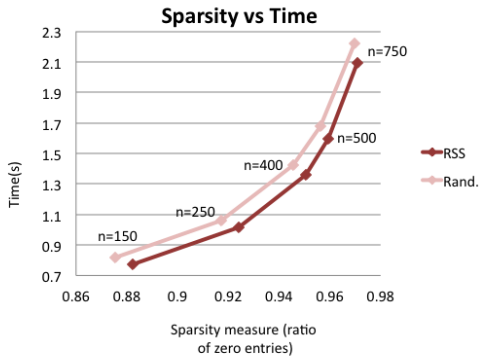
(a) Graph 1



(b) Graph 2



(c) Graph 3



(d) Graph 4

図 4　Graphs demonstrating the results.

In comparing the results using our RSS method against the random selection method, we can draw two conclusions. First, our method outperforms the random approach in terms of accuracy for each value of n. Moreover, the difference becomes more apparent as the sensing matrix becomes smaller in dimension, as illustrated in Graph 1. This observation coincides with the work in [13]. Another important observation to make is that the processing time is always slightly slower using the random approach, which implies that the RSS approach increases the incoherence between different classes in the training data, thereby increasing the sparsity of the recovered signals and improving the time efficiency. As stated in [3], sparsity of a signal directly determines how efficient it can be recovered. These observations are illustrated in Graph 2 and Graph 4.

4.3 Discussion

Our method yields several benefits and possibilities for real life face recognition systems. Assuming we have enough test data to obtain the necessary l1-minimization output. The information can then be used to effectively reduce n and improve the quality of the data simultaneously by neglecting unrepresentative items that are likely to be noisy.

While method based on joint sparsity models [4] can also perform face recognition with reasonable accuracy by using an even smaller set of training data, such approach removes information of the individual signal in relation to the training data. For instance, we can use ranked list of coefficients generated in RSS for a test sample to get the most similar items within the same class in addition to the identification of the class.

From 図 5, we can see that our RSS method provides the distribution of the representativeness in each class by capturing the variability in the training set. As shown in the example, the top ranked items roughly covers the different poses with little occlusion while items with more corruption or occlusion are likely to become outliers, therefore ranked at the bottom of the list.

5. Conclusion and Future Work

We have proposed a new approach for SRC that utilizes the output of l1-minimization for a set of validated samples to effectively select representative items in the training set. By reducing the number of columns in the sensing matrix A, RSS is able to greatly improve the space and time efficiency of the recognition task without losing significant accuracy. It has also been shown to be more accurate and efficient than a random selection approach.

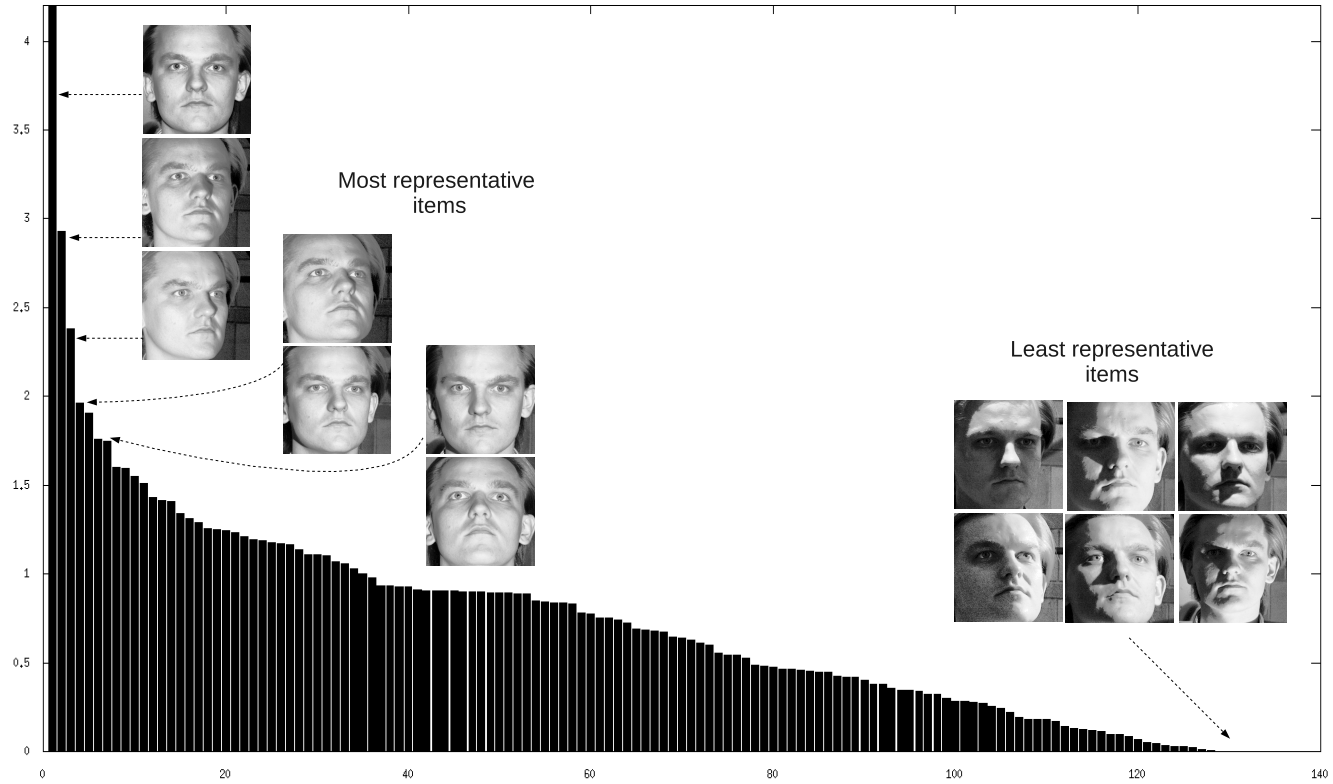


図 5 This graph contains all the training samples in one of the classes, ranked by the representativeness of each item based on its normalized score.

One idea for a future application is to use RSS as a method to incrementally update a set of training data by incorporating the score of each coefficient associated with each item in a newly recovered signal and the original calculated score. Such approach may be able to enhance the representativeness, hence quality of a training set and adapt to changes in a class over time. The scoring scheme may also be applied to validate new input or to determine if a new sample should become a new representative item in the class.

Some probable improvements include finding ways to effectively determine the optimal number of representative samples required to represent a class, which is dependent on the variability of the data. It may also be interesting to apply our method to more diverse or noisy data such as images from the web.

文　　献

[1] H. Bay. Surf: Speeded-up robust features. In *Computer Vision and Image Understanding*, June 2008.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] E.J. Candes and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, Vol. 25, No. 2, pp. 21–31, March 2008.

[4] S.F. Cotter. Sparse representation for accurate classification of corrupted and occluded facial expressions. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010.

[5] D.L. Donoho. Compressed sensing. In *IEEE Trans. Inform. Theory*, July 2006.

[6] M. Figueiredo. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. In *IEEE Journal on Selected Topics in Signal Processing*, 2004.

[7] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, Vol. 23, No. 6, pp. 643–660, 2001.

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. In *Int. Journal of Computer Vision*, January 2004.

[9] D.M. Malioutov. Homotopy continuation for sparse signal representation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2005.

[10] P. Nagesh. A compressive sensing approach for expression-invariant face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2009.

[11] M. Turk. Eigenfaces for face recognition. In *Journal of Cognitive Neuroscience*, 1991.

[12] J. Wright. Robust face recognition via sparse representation. In *IEEE Trans. PAMI*, February 2009.

[13] A. Yang. Feature selection in face recognition: A sparse representation perspective. In *UC Berkeley Technical Report UCB/EECS-2007-99*, August 2007.

[14] A. Yang. Fast l1-minimization algorithms and an application in robust face recognition: A review. In *UC Berkeley Technical Report UCB/EECS-2010-13*, 2010.