Dictionary Construction for Sparse Representation Classification: A Novel Cluster-based Approach

Weiyang Liu[†], Yandong Wen[‡], Hui Li[†], and Bing Zhu[†]
[†]School of Electronic & Computer Engineering, Shenzhen Graduate School, Peking University, P. R. China
[‡]School of Electronic & Information Engineering, South China University of Technology, P. R. China
E-mail: wyliu@pku.edu.cn, wen.yandong@mail.scut.edu.cn, lih64@pkusz.edu.cn, zhubing@sz.pku.edu.cn

Abstract— **There has been a rapid development in sparse representation classification (SRC) since it came out. Most previous work on dictionary improvement was to enhance the classification performance by modifying the dictionary representation structure while this paper concentrates on the reduction of dictionary length with nearly no sacrifice in classification accuracy. A novel cluster-based dictionary construction approach for SRC is proposed in this paper. Both cluster technique and clustering evaluation index are introduced to help construct an optimal dictionary for better classification performance. Results of experiments have verified that the new dictionary does not lose discrimination ability while its running time is greatly reduced. Most importantly, its robustness is also preserved.**

Keywords— Dictionary Construction; Sparse Representation Classification; Cluster Technique; Optimization;

I. INTRODUCTION

Sparse Representation Classification (SRC) [1] serves as an emerging and powerful technique in pattern recognition, which takes advantage of the sparsity to represent a signal with only a few non-zero elements, enabling it to be classified by the reconstruction residual errors. Huang et al. [2] sparsely coded a signal over a set of redundant basis and classified the signal according to the coding vector. John Wright et al. [1] reported a very interesting research about applying sparse representation to robust face recognition and comprehensively introduced the novel recognition method, namely SRC. SRC provides us with a brand new framework to classify signals, which leads to the emergence of various applications [1,3,4].

With large dictionary, SRC framework performs poorly in real-time situation for its high computational complexity, especially when the l_1 norm minimization algorithm is applied. Therefore, it is necessary to enhance the real-time performance of SRC for practical use. Generally speaking, there are two fundamental ways to enhance the performance of SRC. To be specific, one method is to use an efficient reconstruction algorithm such as Tree Orthogonal Matching Pursuit (TOMP) algorithm [5] and Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [6] in place of the Basis Pursuit (BP) algorithm [7]. This kind of performance-enhancing methods, such as TOMP algorithm, may sacrifice too much recognition accuracy. The other method is to modify the dictionary directly. Michal Aharon and Micheal Elad came up with an algorithm for designing overcomplete dictionaries for sparse representation (K-SVD) [8]. Zhang et al. [9] successfully applied K-SVD algorithm to construct a discriminative dictionary. Their work focused on the structure optimization of dictionary. However, our approach is to optimize the dictionary from the perspective of atom representation and dictionary length (number of atoms). That is to say, we concentrate on reducing the dictionary length without losing any classification accuracy. But why is it possible to achieve? And how to achieve it?

An ideal dictionary in SRC is constituted by face samples which are widely distributed within each subject and share no similarities between arbitrary two different subjects. Unfortunately, the reality is that when face samples in each subject are widely distributed and plenty enough, it is very likely that misclassification happens, which inevitably harms the classification accuracy. As a result, cluster technique with cluster evaluation index is adopted to optimally construct new atoms to represent the original subject for better classification accuracy and lower computational complexity.

In general scenario, there is no exact rule for selecting training images, which may lead to too much redundancy of each subject in training images. Moreover, too many training images might cause the overfitting problem in recognition tasks. Therefore, our approach can be smoothly applied here to optimally reduce the dictionary length to a proper level and still maintain the original discrimination ability. Moreover, the classification efficiency can also benefit from the optimized dictionary. In other word, the running time of classification is greatly reduced thanks to the optimized dictionary.

On the other hand, if the dictionary length is fixed, which indicates that the computational complexity is equal, the clustered dictionary has more sample information and less between-subject interference compared to the shortened dictionary randomly constituted by raw atoms (atoms that have not been clustered), leading to stronger discrimination ability.

Highly motivated by the need of dictionary construction, this paper explores to find a proper way to construct a better dictionary for SRC. Inspired by the recent K-SVD algorithm, we smoothly combine the clustering technique as well as the clustering evaluation method and propose a novel cluster-based

dictionary construction model for SRC framework. The outline of this paper is as follows. In section 2, we briefly introduce the SRC framework. In section 3, a cluster-based dictionary construction model is established and discussed. Section 4 presents the experiments performed via Matlab, followed by brief discussions and concluding remarks in Section 5.

II. SPARSE REPRESENTATION CLASSIFICATION FRAMEWORK

Assume there exist i sample subjects which contain n_i training samples respectively. Then each subject can be written as $D_i = [v_{i1}, v_{i2}, \cdots, v_{in_i}] \in \mathrm{R}^{m \times n_i}, i \geq 1$. Assume the size of each training image is $w \times h$. To simplify, the training image we study is gray-scale so there exist a matrix that stands for each training image. $v_{ij} \in \mathrm{R}^m (m = wh)$ is a vector, namely atom, constructed by the image matrix. A m length vector representing the image is given by stacking its columns.

An input test image belonging to the ith subject can be expressed as the following equation:

$$y = \alpha_{i1} v_{i1} + \alpha_{i2} v_{i2} + \cdots + \alpha_{in_i} v_{in_i} \tag{1}$$

where $\alpha_{ij} \in \mathrm{R}, j = 1, 2, \cdots, n_i$.

The equation above applies to the situation when we are aware of the subject that the test image belongs to. But the reality is that the subject of the test image is unknown.

The matrix D_0 is defined to represent all the training images. Assume there are k subjects of training images and every subject has n training images. So D_0 can be written as

$$D_0 = [D_1, D_2, \cdots, D_k]. \tag{2}$$

According to the expression of all the training samples, the test sample y can be expressed as

$$y = D_0 x_0 \in \mathrm{R}^m. \tag{3}$$

where x_0 is a coefficient vector and meanwhile the ideal form of x_0 is $[0, \cdots, 0, \alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{in_i}, 0, \cdots, 0]^T \in \mathrm{R}^n$. Looking into the coefficient vector, we can see that only the coefficients related to the subject that y belongs to are non-zero.

The key to this image recognition problem is $y = D_0 x_0$. By getting the sparsest solution of this equation, we can solve the image recognition problem. This image recognition problem is in fact a l_0 norm minimization problem which is known as a NP-Hard problem [10].

However, it is equivalent to the l_1 norm minimization problem shown as follows, which is solvable by linear programming methods under certain conditions [11,12].

$$\begin{aligned} &\hat{x}_1 = \arg\min \|x\|_1 \\ &\text{subj. to } y = D_0 x \end{aligned} \tag{4}$$

Next is to calculate the residual of each subject and select the smallest one as the classification output. To accomplish that, we need to define a subject-related function $\delta_i(\bullet): \mathbb{R}^n \to \mathbb{R}^n$ first:

$$\delta_i(x) = [0, \cdots, 0, \cdots, \alpha_{i1}, \cdots, \alpha_{in}, 0, \cdots, 0] \tag{5}$$

where $x = [\alpha_{11}, \cdots, \alpha_{1n}, \cdots, \alpha_{i1}, \cdots, \alpha_{in}, \cdots, \alpha_{kn}]$. Using $\delta_i(\bullet)$, we can obtain the reconstructed image $\hat{y}_i = D_0 \delta_i(\hat{x}_1)$. Then we choose the smallest residual as the recognition result:

$$\min_i err_i(y) = \left\| y - D_0 \delta_i(\hat{x}_1) \right\|_2 \tag{6}$$

where residual err is defined as 2-norm of the difference between the test image y and the reconstructed image $\hat{y}$.

III. CLUSTER-BASED DICTIONARY CONSTRUCTION MODEL

A. Necessity and Feasibility of Cluster-based Dictionary Optimization

SRC requires a dictionary with enough within-subject training images, which makes the dictionary too large. And such large dictionary leads to two problems. One is the between-subject interference and another is the high computational complexity. However, as we know, the dictionary used in SRC has not been processed, or in other word, been optimized. Optimized dictionary is constructed by new atoms clustered from the original dictionary, decreasing its redundancy and between-subject interference. This clustered dictionary tremendously reduces the time of recognition while preserving the recognition rate compared to the original dictionary. As a result, it is necessary to optimize the dictionary in order to enhance its recognition accuracy and real-time performance.

The feasibility of dictionary optimization lies in the pre-experiment shown below. To be simple, the purpose is to determine whether a dictionary can be optimized from the perspective of dictionary length. Dictionary optimization here is to reduce the length of the original dictionary to an optimal level.

To study the recognition ability with fewer atoms, we do a simple experiment. With all training samples and test samples, it is possible to obtain the best way to exclude atoms in the dictionary. Suppose the newly-constructed dictionary is reduced to r atoms and the original dictionary has g atoms. According to the idea of exhaustion, we know there exist C_g^{g-r} kinds of way to obtain a new dictionary with r raw atoms. If we run the recognition test with all given test samples (test samples stay unchanged during the whole experiment) over the dictionary, then we can get the recognition rate (the number of successful recognition divided by the number of recognition tests) of this dictionary. Theoretically Speaking, the dictionary that has the highest recognition rate of all the C_g^{g-r} dictionaries is the best dictionary constructed from r raw atoms. In the light of it, it is possible to obtain a relation between the number of atoms and the recognition rate. Because of the extremely large complexity of the exhaustion calculation, we use the random test (this test is to randomly abandon $g-r$ atoms each

time, which is repeated 200 times to select the dictionary of the highest recognition rate) to estimate the relation.

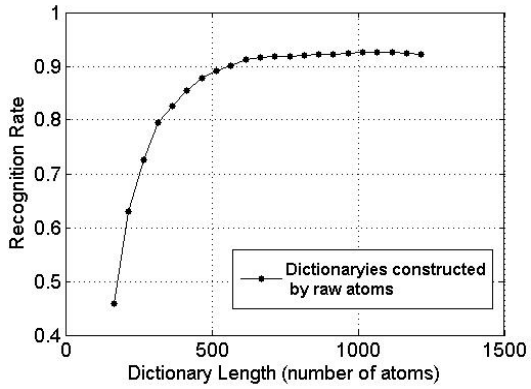

Figure 1. Relation between number of atoms and optimal recognition rate. The experiment is run in extended YaleB face database (38 individuals), and the original dictionary contains 38 subjects. In each subject, 32 images are randomly selected from database for training. The dimension of each atom is downsampled to 176 (16×11).

From Fig.1, it can be seen that there is lots of potential to compress the original dictionary without losing too much recognition rate. However, in practical use, it is impossible to get all the test samples in advance because most test samples are unknown. The relation we discovered can not be calculated in practical use, and additionally, the curve shown in Fig.1 is close to the upper boundary of dictionaries constructed by raw atoms but is not the theoretical best recognition rate of the same atom number. It is indicated that the approach we used to construct the dictionary can surpass this curve. So the relation between number of atoms and optimal recognition rate is just a baseline for us. So this paper explores to find an approach to construct the optimal dictionary by utilizing the dictionary structure, because there are no enough test samples to find out whether a dictionary is good or not in practical. The recognition rate of the new dictionary should be close to or higher than the recognition rate in Fig.1 without any prior test samples.

B. General Model for Cluster-based Dictionary Optimization

Our approach to optimize the dictionary, simply speaking, is to cluster atoms of the dictionary using certain clustering algorithms. The general model includes three necessary parts. First is to choose a representation domain (feature extraction) to do the clustering. Second is to construct a clustering evaluation index. The last is to apply a clustering algorithm. This approach can be summarized as a general model.

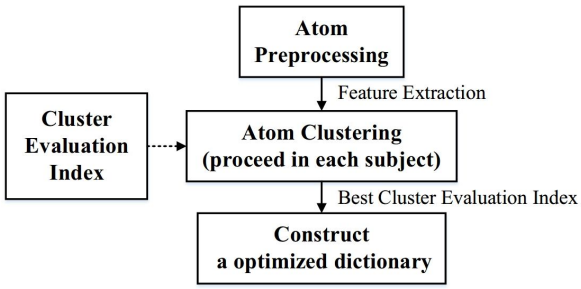

Figure 2. General Model for Cluster-based Dictionary Optimization

The clustering evaluation index, denoted by $\mathrm{CEI}(\bullet)$, is constructed to quantify the sample representation ability of a certain cluster choice. Before discussing the optimization problem, we need to define the clustering granularity. This optimization is, in fact, run in each subject of the dictionary, but here we assume there only exist one subject in the dictionary in order to simplify the theoretical analysis.

Definition 1. *(granularity)*: The scatter degree of a clustering choice is defined as *granularity*, namely

$$\mathrm{Gran}(\Omega)=\frac{CD(\Omega)}{MCD} \tag{7}$$

where Ω is one possible choice of the clustering results, $CD(\bullet)$ is the cluster distance of the clustering choice and maximum cluster distance is denoted by MCD.

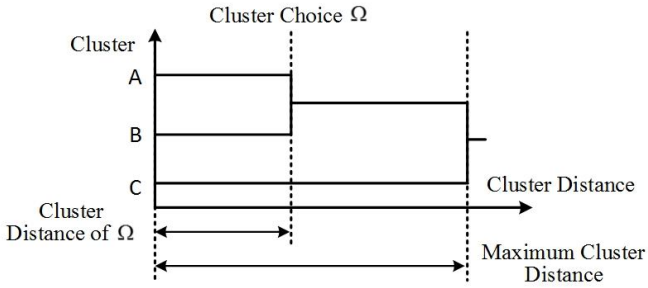

Figure 3. Illustration of Cluster Distance and Maximum Cluster Distance

Definition 2. *(the partition of dictionary)*: Suppose all the atoms in the dictionary $D=\{v_i\}, i\in\Gamma$ are indexed by a set Γ, $D_\Lambda=\{v_i\}, i\in\Lambda$ is defined as sub-dictionary with $\Lambda\subset\Gamma$ and $\Lambda\neq\varnothing$ satisfied. The sub-dictionary set $\{D_{\Lambda_i}, 1\le i\le n\}$ is defined as the partition of the dictionary if $\bigcup_{i=1}^{n}\Lambda_i=\Gamma$ and $\Lambda_i\bigcap\Lambda_j=\varnothing, \forall i\neq j$.

Definition 3. *(the smallest similarity)*: The smallest similarity among atoms in sub-dictionary D_Λ, denoted by λ_Λ, is defined as the following formula where $\mathrm{dis}(v_i,v_j)$ is the distance between normalized atom v_i and v_j from D_Λ:

$$\lambda_\Lambda=\min_{i,j\in\Lambda}\left|1-\mathrm{dis}(v_i,v_j)\right| \tag{8}$$

If λ_Λ can be large enough, the dictionary can be represented by a molecule and the optimal molecule that can best represent the dictionary D_Λ can be found by

$$m_\Lambda^{opt}=\arg\min_{m,\ \|m\|_2=1}\sum_{i\in\Lambda}\mathrm{dis}(m,v_i) \tag{9}$$

The cluster-based dictionary optimization can also be summarized as an optimization process formulized as follows:

$$\max_{\mathrm{Gran}(\{D_{\Lambda_i},1\le i\le n\})} \mathrm{CEI}\left(\{D_{\Lambda_i},1\le \mathrm{i}\le \mathrm{n}\}\right)$$
$$\text{subj. to } \begin{cases} m_{\Lambda_i}^{opt} \in [\mathrm{D}] \\ \bigcup_{i=1}^{n} D_{\Lambda_i} = D \\ 1 \le n \le \dim(D) \end{cases} \tag{10}$$

where $\{D_{v_i},1\le \mathrm{i}\le \mathrm{n}\}$ stands for the clustering result and also is the subset of the linear span of D, namely $[\mathrm{D}]$. $\dim(\cdot)$ is the dimension (number of atoms) of samples in a dictionary. $\mathrm{Gran}(\bullet)$ means traversal of all possible outcomes of the clustering algorithm.

In this paper, we will present a specific case of the general cluster-based dictionary optimization model by using hierarchical cluster tree and SD index [13]. However, the method presented here is not the only way to construct an optimal dictionary but just one of the many. The hierarchical cluster tree is just one of the many clustering methods while SD index is also one of the many clustering evaluation indexes.

C. *Dictionary Optimization via Hierarchical Cluster Tree and SD Index*

Hierarchical cluster tree [14] is applied to cluster the atoms (All atoms are normalized in advance.) in the original dictionary and therefore molecules are generated to represent similar atoms. Before the hierarchical cluster tree is established, the distance between two atoms is defined as follows:

$$\mathrm{dis}(v_i, v_j) = 1 - \left|\langle v_i, v_j \rangle\right| \tag{11}$$

where, v_i and v_i are normalized atoms. Moreover, there are also some other available distances that can be used such as Euclidean distance and correlation distance.

Plugging (12) into (10), we can obtain

$$\begin{aligned} m_{\Lambda}^{opt} &= \arg\min_{m,\ \|m\|_2=1} \sum_{i\in\Lambda} \left(1 - \left|\langle m, g_i \rangle\right|^2\right) \\ &= \arg\max_{m,\ \|m\|_2=1} \sum_{i\in\Lambda} \left|\langle m, g_i \rangle\right|^2 \\ &= \arg\max_{m,\ \|m\|_2=1} \sum_{i\in\Lambda} m^* A_{\Lambda} A_{\Lambda}^* m \end{aligned} \tag{12}$$

where every row of A_{Λ} is an atom in the sub-dictionary D_{Λ}. m_{Λ}^{opt} is the same as the eigenvector of $A_{\Lambda}^* A_{\Lambda}$ with the maximum eigenvalue.

We generate the hierarchical cluster tree from atoms of the same subject. So if there are n subjects in the dictionary, there will also be n hierarchical cluster trees. In short, we need to reduce the size of the original dictionary by selecting the most representative molecules. The following figure is to show how a hierarchical cluster tress is built.

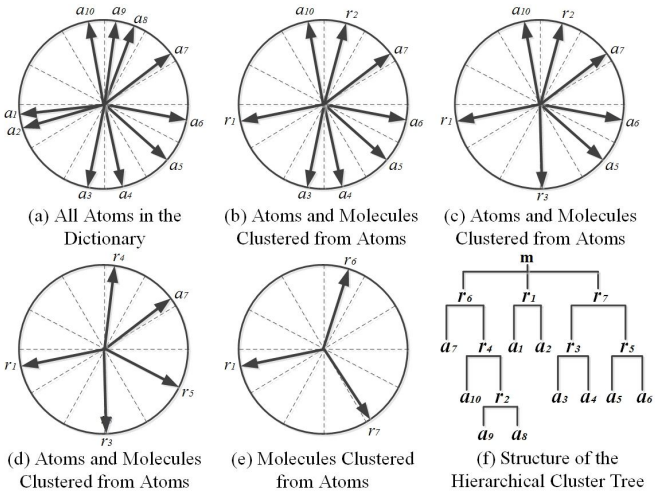

Figure 4. Construction Process of the Hierarchical Cluster Tree

The hierarchical cluster tree is generated by certain mechanism. First, every atom has the same weight and this algorithm selects two closest atoms to construct a new atom, namely a molecule whose weight equals two atoms. And the distance is weighted. This operation keeps repeating until all the atoms become one molecule just like Fig.4(f).

The $\mathrm{CEI}(\bullet)$ applied here is SD-index [13] which can objectively evaluate the clustering scheme (choice of the partition) regardless of different clustering algorithms. SD-index is based on the validity index defined for FCM in [15]. Let v_i be the molecule that optimally represent the ith cluster (v_i also called cluster center). $\sigma(v_i)$ and $\sigma(X)$ stand for the variance of the ith cluster and the whole subject respectively. The atom distance used in [13] is replaced with the distance defined above. Average scattering for cluster is defined as

$$\mathrm{Scat}(c) = \frac{1}{c}\sum_{i=1}^{c} \frac{\left\|\sigma(v_i)\right\|_2}{\left\|\sigma(X)\right\|_2} \tag{13}$$

where c is the number of clusters. Total separation between clusters is defined as

$$\mathrm{Dist(c)} = \frac{d_{\max}}{d_{\min}} \sum_{k=1}^{c} \left(\sum_{z=1}^{c} \mathrm{dis}(v_k, v_z) \right)^{-1} \tag{14}$$

in which $d_{\max} = \max\left(\mathrm{dis}(v_i, v_j)\right)\ \forall i, j \in \{1,2,\cdots,c\}$ is the maximum distance between cluster centers and $d_{\min} = \min\left(\mathrm{dis}(v_i, v_j)\right)$ is the minimum distance between cluster centers. After combination of (13) and (14), SD-index is defined as

$$\mathrm{SD}(c) = \alpha\,\mathrm{Scat}(c) + \mathrm{Dist}(c) \tag{15}$$

where α is a weighting factor usually equal to $\mathrm{Dist}(c_{\max})$. $c_{\max}$ is the maximum number of input clusters. The first term $Scat(c)$ indicates the average compactness of clusters (intra-cluster distance) and the second term $\mathrm{Dist}(c)$ indicates the sep-

aration between the c clusters (inter-cluster distance). We apply the specific model to each subject of the dictionary and construct a new optimized dictionary for SRC.

IV. EXPERIMENTS AND RESULTS

Experiments in the section are performed in the extended Yale B database [16] (38 individuals, 2414 images) and the AR database [17] (126 individuals, over 4000 images). Each image in extended Yale B database is downsampled to 176 dimensions (16×11) and each image in AR database is also downsampled to 1260 dimensions (42×30). We do the classification performance test over extended Yale B database. The original dictionary (1216 atoms) is constructed by 38 subjects. In each subject, there are 32 images randomly selected from the database for training and the left images (1198 test faces) are for testing. For AR database, we use it to simulate the robustness test. In AR database, 26 pictures were taken in two separate sessions [1]. These images include more facial variations (illumination change, expressions and facial disguises). We chose a subset of the database consisting of 50 males and 50 females. We use the 1400 images (100 individuals) with only illumination and expressions change as our training samples and all the images with glass occlusion (600 images, 100 individuals) as our test samples, which is similar to the experiment setup in [1]. In fact, experiments in this section are all repeated 20 times and use the average results.

Three experiments are designed. First is to simulate the cluster-based dictionary optimization model without clustering evaluation index. We traverse Gran(•) and obtain a new dictionary to run the recognition test whose result is shown in Fig.6. The second experiment is to simulate the dictionary optimization based on hierarchical cluster tree and SD index and the result is shown in Table 1. The first two experiments are run in Extended Yale B database. The last experiment shown in Fig.7 is the robustness test using AR face database. For better illustration, an example of cluster-based dictionary optimization is shown in Fig.5.

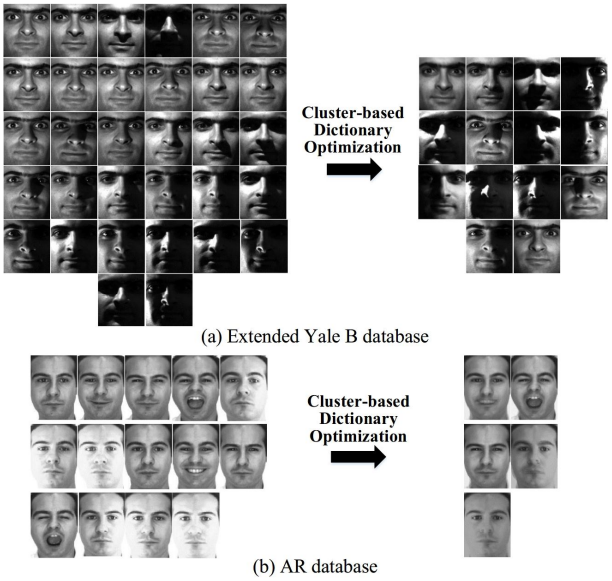

Figure 5. (a) Extended Yale B database. Faces on left are from the original dictionary constructed by faces of the eleventh subject in Yale B database, while faces on right are generated via our approach. (b) AR database. Faces on left are from the original dictionary constructed by faces of the first subject in AR database, while faces on right are generated via our approach.

TABLE 1. RESULT OF DICTIONARY OPTIMIZATION BASED ON HIERARCHICAL CLUSTER TREE AND SD INDEX.*

	Dictionary Length	Recognition Rate	Running Time (s)
Optimized Dictionary (Clustered Atoms)	**588.2 (±10.8)**	**0.9154 (±0.0220)**	**89.71 (±3.53)**
Original Dictionary	**1216**	**0.9232 (±0.0331)**	**181.51 (±7.91)**
Shortened Dictionary (Raw Atoms)	**590**	**0.8825 (±0.0291)**	**90.23 (±3.87)**

*The simulation is repeated 20 times in extended Yale B database and the result is averaged. Running time experiments are tested in a computer with 3.06GHz dual-core CPU and 8G RAM.

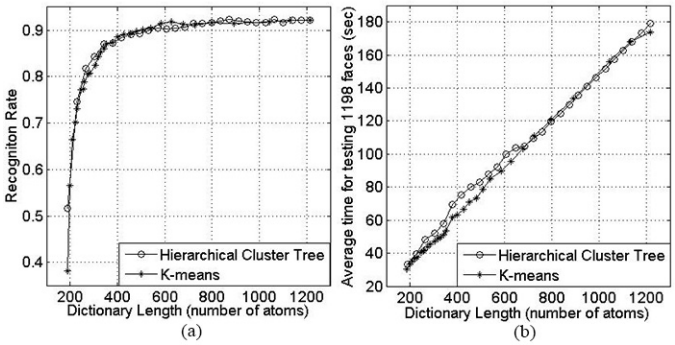

Figure 6. (a) Relation between recognition rate and dictionary length. Two clustering method are used here while K-means [18] is for comparison. In K-means, we assign the number of clusters to every subject to control the dictionary length while in hierarchical cluster tree we use granularity to control the dictionary length. (b) Relation between computation time and dictionary length. Its simulation environment is the same as (a).

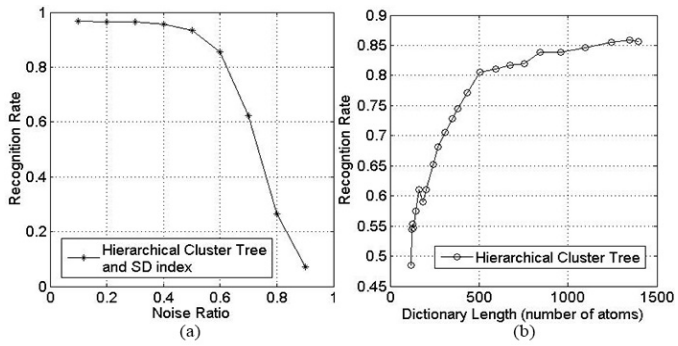

Figure 7. (a) Relation between noise ratio and recognition rate. The noise robustness simulation aims at the model establish in section 3.3. (b) Relation between dictionary length and recognition rate in glass occlusion. This experiment is performed in AR face database. In the category of glass occlusion situation, we randomly use 1300 test samples and 1300 training samples as dictionary. The result averages 20 times experiments.

In the example of cluster-based dictionary optimization shown in Fig.5, we can see that the redundancy of the optimized dictionary is reduced while the discriminative information of the original dictionary is still preserved. For the first experimental result shown in Fig.6(a), it can be concluded that the cluster-based dictionary optimization method has the feasi-

bility to reduce the length of the original dictionary and meanwhile maintain the discrimination ability. Fig.6(b) indicates that there is strong linear relationship between the computational complexity (running time) and the number of atoms (dictionary length). Fig.7 shows that the robustness of the optimized dictionary is preserved. Applying the proposed specific dictionary optimization method, we obtain the results in Table.1 which proves the optimized dictionary has the similar recognition rate with the original dictionary, but its running time is reduced to less than a half.

V. CONCLUSIONS

In this paper, we have proposed a novel cluster-based dictionary construction model for SRC framework. Based on a clustering process, this model explores to construct an optimal dictionary with fewer atoms. As an instantiation of the general model, a specific dictionary construction model based on hierarchical cluster tree and SD index is also presented and comprehensively discussed in this paper. We demonstrate that SRC framework benefits from the cluster-based dictionary optimization for the reduction of computation time, which is also validated by the results of experiments. Moreover, simulations have verified that the discrimination ability and robustness are both preserved.

ACKNOWLEDGMENT

The first author acknowledges Prof. Zhiheng Zhou for motivating the sparse representation work and appreciates the help of Ms. Lijia Lu for proofreading. Our work is supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB315904, the National Natural Science Foundation of China under Grant NSFC61179028, the Natural Science Foundation of Guangdong Province under Grant NSFGD S2013020012822, the Basic Research of Shenzhen under Grant SZ JCYJ20130331144502026.

REFERENCES

[1] Wright John et al. "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.31, pp.210-227, 2009.

[2] Huang Ke and Selin Aviyente, "Sparse representation for signal classification,"*Advances in neural information processing systems*, 2006.

[3] Wan Vincent and Steve Renals, "Speaker verification using sequence discriminant support vector machines,"*Speech and Audio Processing, IEEE Transactions on*, vol.13, pp.203-210, 2005.

[4] Chen Yi, Nasser M. Nasrabadi, and Trac D. Tran, "Hyperspectral image classification using dictionary-based sparse representation,"*Geoscience and Remote Sensing, IEEE Transactions on*, vol.49, pp.3973-3985, 2011.

[5] La Chinh and Minh N. Do, "Signal reconstruction using sparse tree representations,"*Optics & Photonics 2005, International Society for Optics and Photonics*, 2005.

[6] Beck, Amir, and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM Journal on Imaging Sciences* 2.1: 183-202, 2009.

[7] Chen, Scott Shaobing, David L. Donoho, and Michael A. Saunders. "Atomic decomposition by basis pursuit." *SIAM journal on scientific computing* 20.1: 33-61, 1998.

[8] Michal Aharon, Michael Elad and Alfred Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,"*Signal Processing, IEEE Transactions on*, vol.54, pp.4311-4322, 2006.

[9] Zhang, Qiang, and Baoxin Li. "Discriminative k-svd for dictionary learning in face recognition." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[10] Rachlin Yaron and Dror Baron, "The secrecy of compressed sensing measurements,"*Communication, Control, and Computing, 2008 46th Annual Allerton Conference on, IEEE*, 2008.

[11] Donoho David L, "Compressed sensing." *Information Theory, IEEE Transactions on*, vol.52, pp.1289-1306, 2006.

[12] Candès, Emmanuel J., Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,"*Information Theory, IEEE Transactions on*, vol.52, pp.489-509, 2006.

[13] Halkidi Maria, Michalis Vazirgiannis, and Yannis Batistakis, "Quality scheme assessment in the clustering process." *Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, pp.265-276, 2000.

[14] Farenzena, Michela, Andrea Fusiello, and Riccardo Gherardi. "Structure-and-motion pipeline on a hierarchical cluster tree." *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009.

[15] Fayyad Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From data mining to knowledge discovery in databases,"*AI magazine*, vol.17, pp.37, 1996.

[16] Georghiades, Athinodoros S., Peter N. Belhumeur, and David J. Kriegman. "From few to many: Illumination cone models for face recognition under variable lighting and pose." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* vol.23.6: pp.643-660, 2001.

[17] Martinez, Aleix M. "The AR face database." *CVC Technical Report* 24, 1998.

[18] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* vol.28.1: pp.100-108, 1979.