Learning of Graph Compressed Dictionaries for Sparse Representation Classification

Farshad Nourbakhsh and Eric Granger

Laboratoire d'imagerie de vision et d'intelligence artificielle

École de technologie supérieure, Université du Québec, Montréal, Canada

fnourbakhsh@livia.etsmtl.ca, Eric.Granger@etsmtl.ca

Abstract: Despite the limited target data available to design face models in video surveillance applications, many faces of non-target individuals may be captured in operational environments, and over multiple cameras, to improve robustness to variations. This paper focuses on Sparse Representation Classification (SRC) techniques that are suitable for the design of still-to-video FR systems based on under-sampled dictionaries. The limited reference data available during enrolment is complemented by an over-complete external dictionary that is formed with an abundance of faces from non-target individuals. In this paper, the Graph-Compressed Dictionary Learning (GCDL) technique is proposed to learn compact auxiliary dictionaries for SRC. GCDL is based on matrix factorization, and allows to maintain a high level of accuracy with compressed dictionaries because it exploits structural information to represent intra-class variations. Graph factorization compression has been shown to efficiently compress data, and can therefore rapidly construct compressed dictionaries. Accuracy and efficiency of the proposed technique is assessed and compared to reference sparse coding and dictionary learning technique using videos from the CAS-PEAL database. GCDL is shown to provide fast matching and adaptation of compressed dictionaries to new reference faces from the video surveillance environments.

1 INTRODUCTION

In the recent years, sparse modelling has become an important tool in the pattern recognition and computer vision communities. They have been successfully applied in many image/video processing tasks like face recognition, image denoising and super-resolution (Mairal et al., 2014). The well-known Sparse Representation Classification (SRC) techniques (Wright et al., 2009) typically need a sufficient amount of representative training data to construct over-complete dictionaries that can provide a high level of performance. In many real-world applications, the number of reference images that are available for system design is limited. Classification systems designed with few reference samples per class are less robust to the intra-class variabilities encountered during operations.

In video surveillance, the amount of reference stills and videos captured during enrolment to design a face recognition (FR) system is typically limited, especially in watch-list screening applications (Dewan et al., 2016). The still-to-video FR systems employed for watch-list screening seek to match probe face images captured using surveillance cameras against the reference still images of each individual of interest enrolled in the gallery. The appearance of faces captured under uncontrolled conditions and over a network of surveillance cameras, varies dues to changes in illumination, pose, expression, scale, blur, etc., and create discrepancies between probe and gallery images.

To enhance the robustness of facial models to intra-class variability, the methods proposed for Single Sample Per Person (SSPP) problems (Tan et al., 2006) performs FR using multiple synthetically generated reference faces from the original reference image (Mokhayeri et al., 2015), multiple face representation (Bashbaghi et al., 2014) and utilizing external data captured from non-target individuals in the operational environment.

In video surveillance applications, many faces of non-target individuals may be captured in the operational environment, and over multiple cameras, to improve robustness to intra-class variations. This paper focuses on SRC techniques based on under-sampled dictionary that are appropriate for the design of still-

to-video FR systems. The lack of reference facial samples available to design face models is compensated by using an over-complete external dictionary that is formed based on non-target individual captured.

Many techniques in SRC are developed to utilize such external data. For instance, Extended SRC (Deng et al., 2012) was one of the first methods that modified SRC framework to achieve benefit of external data on under-sampled dictionary. RADL and SVDL (Wei and Wang, 2015; Yang et al., 2013) are more recent techniques on sparse modelling that combine dictionary learning and classification to manage external data more efficiently.

One of the main drawbacks of using large external data is storage and computational time. The time and memory complexity is proportional to the size of dictionary $|D|$. Dictionary learning techniques are a possible solution to build a compact representation of the large external data (Mairal et al., 2014). These have been used by researchers on SRC in recent years, and can be categorised into structural, hierarchical and topological techniques. Dictionary learning methods have been reviewed and compared in literature (Shafiee et al., 2013). For instance, K-Means Singular Value Decomposition (K-SVD) that is a common method in literature (Aharon et al., 2006) to learn dictionary from external data. It is worth to mention that although dictionary learning techniques provides a compact representation of data, they are not necessarily time efficient. It has been shown that learning a dictionary based on sparse representation is NP-hard (Tillmann, 2015).

One way to reduce memory and time complexity, and increase accuracy is data compression (Choi and Szpankowski, 2012; Navlakha et al., 2008; Toivonen et al., 2011) as applied in big data applications (Nourbakhsh, 2015). Compressing data consists in changing its representation in a way that requires fewer bits. Lossy or lossless compression are possible depending on the reversibility of encoding. Many graph compression methods have been proposed in literature with a different focuses, e.g., information theoretic approach, partitioning as a regular pairs and summarization of data. A graph compression method is suitable for a preprocessing of SRC methods that provides cluster-based compactness.

Nourbakhsh et al. (Nourbakhsh et al., 2015) have proposed a graph compression method based on matrix factorization that focuses on structural information. Data is presented as a similarity matrix of data objects then it encodes the order of graph data into a compressed graph of order $k \leq n$ in a way to minimize reconstruction error. It has been shown that the proposed data compression method reduces the complexity of many algorithms from n^2 to $(k^2 + n)$ by replacing the original data with its corresponding factorization. Although this method has not been developed for graph clustering, practical results suggest that it provide a potentially powerful preprocessing tools for dictionary learning and SRC.

In this paper, a new graph compression technique based on matrix factorization is proposed to lean compact auxiliary dictionaries for accurate SRC. This techniques, entitled Graph-Compressed Dictionary Learning (GCDL), allows to maintain a high level of accuracy with compressed dictionaries because it exploits structural information to represent intra-class variations. In addition to providing fast matching, these graph factorization compression has been shown to efficiently compress data, and can therefore rapidly construct compressed dictionaries compared to dictionary learning methods in the literature. It is therefore suitable for adaptation of compressed dictionaries to newly-acquired reference faces captured in changing video surveillance environments.

2 Sparse Representation Classification

In sparse representation, probe samples are a sparse linear combination of reference samples. The sparse representation of a signal can be formulated as follows with the assumption that the samples from a single class lie on a liner subspace. Given an over complete matrix $\mathbf{D} = [D_1, D_2, ..., D_l]$ with the l distinct classes and the size of over complete dictionary is $m \times n$ that the size of column is greater than the size of row and usually $n \gg m$. The ith class of Dictionary can be presented as $D_i = [d_{i,1}, d_{i,2}, ..., d_{i,n_i}]$ with a size of $m \times n_i$. The probe sample y with a size $m \times 1$ from the class D_i will lie in the linear span of over complete dictionary as below:

$$y = a_{i,1}d_{i,1} + a_{i,2}d_{i,2} + ... + a_{i,n}d_{i,n_i}, \quad (1)$$

Where a is the linear representation of y on the over complete dictionary, defined as,

$$y = [D_1, D_2, ..., D_l]x. \quad (2)$$

where x is a coefficient vector. The entries of x are almost zero except the ones belong to the ith class.
In the real FR application because of the presence of some factors such as noise, the linear eq (2) is not determined then there is no unique solution so the possible choice to resolve this issue is using sparse representation of probe sample by increasing the value

of l. By making the sparse representation, the natural solution is applying l_0- norm minimization that is a way to determine the coefficients that is equivalent to the number of non-zero components in vector x.

$$\hat{x} = \min_{x} \|x\|_0 \qquad s.t. \qquad y = \mathbf{D}x. \qquad (3)$$

Eq (3) is *NP*-hard and difficult to approximate. By combining compressed sensing (Donoho and Tsaig, 2008) and sparse representation theory, an approximate solution is obtained by replacing the l_0 norm in eq (4) with the l_1 norm:

$$\hat{x} = \min_{x} \|x\|_1 \qquad s.t. \qquad y = \mathbf{D}x. \qquad (4)$$

If the solution of vector x is sparse enough so the solution of eq (4) is equivalent as eq (3). Eq (4) can be solved with different optimization techniques like basis pursuit using linear programming.

Sparse representation classification and its extension have been used a lot in the literature in last years. We have used difference SRC techniques based on (Wei and Wang, 2015) for comparison with our method so we review some of these techniques to complete this paper.

To make eq (4) slightly robust to noise, eq (5) is proposed which is a generalized version of eq (4), which allows for certain degree of noise that is called LASSO , is to find x such that the following objective function is minimized:

$$\min_{x} \|y - \mathbf{D}x\|_2^2 + \lambda \|x\|_1 \qquad (5)$$

where $\lambda > 0$ controls reconstruction error and sparsity which is called scalar regularization parameter. When the sparse vector coefficient is obtained by eq (5) or (4), the query image is recognized to the class its belongs by calculating the distance between the query image and the reconstructed image based on sparse vector coefficient that indicates the class elements as non zero value and almost zero for non classes. The main idea is when the class is recognized correctly, the query image can be reconstructed back linearly with a relevant bases of dictionary.

SRC method needs large amount of training to built over complete dictionary and the training data size has a direct effect on the final performance accuracy. In many real applications like video surveillance, there is no enough data so one of the solutions is to apply non target external data. Deng et. al (Deng et al., 2012) extended SRC method for undersampled dictionary data. In their method, undresampled dictionary $D = [d_1, d_2, ..., d_l]$ (**D** is overcomplete dictionary) has one or few samples for each object so external dictionary from non target data **ED** that covers the known

distortion like illumination to learn intra class variation adds to the undersampled dictionary as follow

$$\min_{x} \|y - [[d_1, d_2, ..., d_l], ED][{}^{x_d}_{x_{ED}}]\|_2^2 + \lambda \|x\|_1 \qquad (6)$$

To assign probe image to the closest class, like SRC, the distance calculated between query image and reconstructed image respect to the nonzero coefficient x_d and extra coefficient x_{ED}.

Robust Sparse Coding (RSC) is proposed in (Yang et al., 2011b) that is a robust face recognition based on SRC. An extra weighting term **W** gets assign to each elements of probe image. The main idea to assign a weight to each pixel is in FR system, pixels from outlier part of image are less informative than central pixels. For instance eyes and noise part of a face have more information than hair part. Wei and Wang (2015a) have proposed a similar frame work to auxiliary dictionary learning to manage external dictionary and classification as follow.

$$\min_{x} \|\mathbf{W}(y - [[d_1, d_2, ..., d_l], ED][{}^{x_d}_{x_{ED}}])\|_2^2 + \lambda \|x\|_1 \qquad (7)$$

Where *ED* is external dictionary when their method is applies only as classification ($RADL_{wo}$) and it is called as learned dictionary ($RADL_wDL$) from overcomplete external data with an optimization algorithm then each column of learned dictionary *ED* is called as **atom**.

3   Dictionary Learning

The performance of SRC methods is limited by the number of reference sample data. For instance, the time complexity respect to the number of sample data is quadratic (**?**). This challenge has been addressed in the literature by applying a compact representation or reducing the number of reference data. For instance, Wright et al. (2009) has suggested a random selection of reference training sample to reduce the time complexity although it impacts the accuracy. The proposed solution in the literature to reduce the time complexity of SRC is to apply Dictionary Learning techniques although most of them are applied offline because of their time complexity.

Olshausen and Field (1996) introduced dictionary learning in the community of pattern recognition. They proposed an unsupervised method based on structure of data that learns the bases of dictionary from training data which is different from classical methods such as discrete Fourier transform (DFT) and various types of Wavelet methods which use fixed

standardized format dictionaries. Dictionary learning can be viewed in different point of views like matrix factorization, risk minimization and constrained variants.

Dictionary learning methods have been applied as preprocessing step for the SRC recently. For example Shafiee et al. (2013) have investigated the effect of three different dictionary learning methods on SRC method. Performance of SRC method affects by number of training data, they used Metaface dictionary learning (Yang et al., 2010b), Fisher Discriminative Dictionary Learning (FDDL) (Yang et al., 2011a), Sparse Modelling Representative Selection (SMRS) (Elhamifar et al., 2012) to obtain compact representation of training data. They were able to show that FDDL method performs a high recognition accuracy compare to other methods although SMRS method requires a much smaller learning time compare to others.

K-Means Singular Value Decomposition (KSVD) (Aharon et al., 2006) and Method of Optimal Directions (MOD) (?) proposed two unsupervised dictionary learning techniques which have been used in the literature. These are EM style methods and alternate between dictionary and sparse coding. The difference between these two methods are in dictionary updating which KSVD updates atom by atom but MOD updates all the atoms simultaneously.

(Ramírez et al., 2010) proposed a frame work to manage sparse modelling and clustering. They introduce set of dictionaries which are optimized one for each cluster. The learned data is a cluster of the union of low dimensional subspaces. Most of the reported dictionary learning methods require an overcomplete dictionary to generalize well the intraclass variations except (Wei and Wang, 2015; Yang et al., 2013) in some extend. However dictionary learning methods based on SR provide a compact representation of overcomplete dictionary that reduces the time complexity of SRCs method, they are almost *NP*-hard to execute so they are used mostly off-line.

4 Graph Compression for SRC

Most of dictionary learning methods for SRC require considerable amount of time to learn dictionary and this construction time increases by the size of input data. In addition, sparse representation methods are *NP*- hard problem. So we need a method to construct dictionary quickly as possible without losing performance. One of the methods deal with big-data/over-complete dictionary is graph compression which is changing the representation of data that re-

quires less memory. Depending on the type of encoding we might have a lossy or lossless compression. Data can be presented as a collection of feature vectors or representation of the similarity/dissimilarity relations among data objects. Therefore, it can be easily convert data to the adjacency matrix of a weighted graph. Compression methods can be addressed with an Information-theoretic approach by compressing graphical structures (Choi and Szpankowski, 2012) without preventing a graph structure as the compressed representation. As a second category can be presented as regularity pairs by Szemerédi regularity lemma (Szemerédi, 1978) that is a well-known result in extremal graph theory, which roughly states that a dense graph can be approximated by a bounded number of random bipartite graphs. An algorithmic version of this lemma has been used for accelerating pairwise clustering (Sperotto and Pelillo, 2007). Finally, a compression method can take to account the structural information from input data, for example (Navlakha et al., 2008) propose a summarization algorithm for unweighted graphs and (Toivonen et al., 2011) suggested a greedy procedure to determine a set of supernodes and superedges that to approximate a weighted graph.

4.1 Graph-Compressed Dictionary Learning (GCDL)

We present our method as shown in Figure 1. We propose to apply graph compression method on external dictionary from non target data to compress the external dictionary to a compact one. Although, using compact representation of dictionary is not new, it is important to apply a method to be executed quickly as a preprocessing step of SRC methods. Because SRC methods are *NP*-hard in their nature, we choose the Homotopy method for the sparse optimization in our work as, its time efficiency suggested in (Yang et al., 2010a).

Figure 1 illustrates schematic presentation of our proposed method. External data is collected from non target data. The first column represents the natural image of its corresponding row. Overcomplete dictionary is constructed by calculating the difference between each image and it corresponding natural image. Similarity graph calculated based on overcomplete dictionary that each image is a node in the graph and weighted edges present the similarity between nodes. In the Similarity Graph' sub-figure, those nodes are marked with circles that are atoms of compressed dictionary. Finally, each probe sample is sparse linear combination of reference sample/gallery and compressed dictionary.
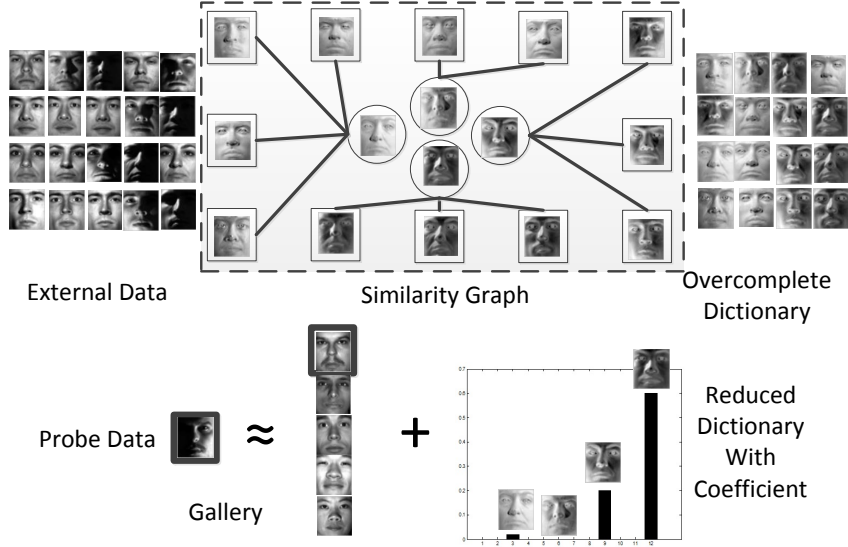
Figure 1: Schematic presentation of Graph Compressed Dictionary Learning (GCDL) method.

Algorithm 1 Graph Compressed Dictionary Learning

Input Data: Over-Complete Dictionary from External Data ED
Output: Graph Compressed Dictionary D_{GCDL}

Graph $G \leftarrow$ Calculate similarity matrix from ED

$X \leftarrow$ draw a random matrix from $\mathcal{S}$
while not stopping criterion met **do**
   $C \leftarrow U_C(X)$ /* Update C */
   $X \leftarrow U_X(X,C)$ /* Update X */
end while
Project X to binary left stochastic matrix

$D_{GCDL} \leftarrow$ Select the representative of each cluster based on the mapping matrix X

The edge weighted graph $G = (V,E,w)$ representing a set of n vertices that each vertices V_i is connected to the vertices V_j with an edge weighted and $E \subseteq V \times V$ is the set of edges. The weight $w(i,j)$ is obtained from the following formula.

$$w(i,j) = exp(\frac{-||ed(i) - ed(j)||_2^2}{\zeta^2}) \qquad (8)$$

where ζ is a positive real number and it is bounded to $(0,1]$, $ed(i)$ is a feature vector from external overcomplete dictionary ED and $||.||_2$ is the Euclidean distance between the two values which gives dissimilarity be-

tween two considered elements. The graph G is complete and undirected with order of n that is the size of external overcomplete dictionary ED. Therefore $w(i,j) = w(j,i)$ for all $(i,j) \in E$. A graph G is called as a symmetric matrix G with order of n. Let $\mathsf{k} \le \mathsf{n}$ be a constant representing that is the number of atoms or the order of compressed new graph C. The rate $\frac{\mathsf{k}}{\mathsf{n}}$ is regarded as the graph compression rate.

We need a many to one mapping $\psi: [\mathsf{n}] \to [\mathsf{k}]$ between vertices of the original graph and compressed graph and we need to determine the compressed graph to be able to reduce the order of graph from n to k. To estimate the mapping function and compressed graph, a least squares approximation can be applied on following minimizer by dropping a left-stochastic constrain to a real matrix X.

$$\begin{aligned} \min \quad & f(X,C) = ||G - X^T C X||_2^2 \\ \text{s.t.} \quad & X \in \mathcal{S}, \quad \mathsf{C} \in \mathbb{R}^{\mathsf{k} \times \mathsf{k}}. \end{aligned} \qquad (9)$$

where

$$\begin{aligned} f(X,C) = & \sum_{(i,j) \in \{1,2,...,N\}} \sum_{k,h \in \{1,2,3,...K\}} \delta_{(k,i) \ne (h,j)} X_{ki} X_{hj} \\ & \times (\mathsf{G}_{ij} - \mathsf{C}_{kh})^2 + \sum_{i \in \{1,2,...,N\}} \sum_{k \in \{1,2,3,...K\}} \mathsf{X}_{ki} (\mathsf{G}_{ii} - \mathsf{C}_{kk})^2. \end{aligned}$$

The optimization can be addressed as a EM method which alternates updates of the variable C and updates of the variable X. The minimization approach converges to a stationary point by updating a decrease of the objective function in every iteration.

Update rule for C.

The update rule for the unconstrained matrix C ($U_C(X)$), is obtained based on the first-order partial derivative of f with respect to C to zero (see (Nourbakhsh et al., 2015), proof of Theorem for details).

Update rule for X.

$U_X(X,C)$ is a multiplicative rule for X that is in a similar manner as the ones suggested in (Lee and Seung, 2001) for Non-negative Matrix Factorization (NMF). We say that X is *Karush-Kuhn-Tucker (KKT-)point* for the following optimization if it satisfies the first-order necessary conditions for the local optimality (details are given in (Nourbakhsh et al., 2015)).

$$\begin{aligned} \min \quad & f_C(X) = f(X,C) \\ \text{s.t.} \quad & X \in \mathcal{Y}_Z \end{aligned} \tag{10}$$

where $C \in \mathbb{R}^{k \times k}$, $Z \in \mathcal{S}$ and $\mathcal{Y}_Z = \{X \in \mathcal{S} : (Z_{ki} = 0) \Rightarrow (X_{ki} = 0)\}$.

Algorithm 1 provides the summary of GCDL approach that works as follows. After calculating weighted graph from overcomplete dictionary that is explained in Figure 1. The minimization optimization starts by a random selection of $X \in \mathcal{S}$ continued by repeatedly alternate between updating C and X with their respective update rules that are $U_C(X)$ and $U_X(X,C)$ respectively until convergence. The stopping criteria are distance between Xs of two consecutive iterations is below a given threshold and maximum number of iterations reaches a threshold. This procedure might converge to a local minima and it guarantees a strict decrease of the objective until a KKT-point is reached. Finally, a discrete solution is obtained through projecting of binary left stochastic matrices by setting to 1 the element having highest value in each column of X and put to 0 the rest.

Although, the above compression method is not designed specificity for clustering, it can converge to a good clustering result to generate the compact representation of input data. From this perspective, the mapping X encodes the clustering result. Then the representative vertices of each cluster based on the mapping X produce the dictionary atoms. The complexity of a matrix vector multiplication reduces from n^2 to $(k^2 + n)$.

5 Experimental Results

The section seeks to assess the effectiveness of the proposed GCDL method compare to state of the art in sparse classification. Several SRC families [1] and dictionary learning methods are used for comparison that are developed recently, SRC (Wright et al., 2009), RSC (Yang et al., 2011b), ESRC (Deng et al., 2012), and *RADL* (Wei and Wang, 2015). Baseline considered methods in the literature can be categorized as methods without external dictionary (SRC and RSC) with external dictionary (such as ESRC and $RADL_{wo}$) and with both dictionary learning and classification like (SVDL and $RADL_wDL$). In addition, we compared the processing time to reconstruct compressed dictionary to GCDL and dictionary learning methods with or without classification part.

The results were performed with the CAS-PEAL database (Gao et al., 2008) that is a large-scale Chinese face database which contains pose variations, expression, accessory, lighting and backgrounds. It covers 1040 individuals (595 males and 445 females). For this experiment, we follow the protocol that has been discussed in the paper of (Wei and Wang, 2015). 100 subjects from the neutral category as gallery images D and their corresponding images from accessory category as testing were selected. The accessory category contains 3 images with hats and 3 images with sunglasses so 600 images are collected for testing in total. The 60 subjects with 6 instances of its accessory category from the rest of 1040 individuals are chosen as external data to build dictionary ED. The pixel-based feature vector is obtained by downsampling the original image to 50×40 pixels.

For the first experiment the performance of GCDL is shown by varying the number of atoms (compression rate) with respect to other methods. The 60 individuals are used as external data with 6 instances for each. In our experiment, the size of external data is increased in each experiment by randomly selecting the data from 60 individuals that starts from 1 to 15 individuals for ESRC and $RADL_{wo}$. In a same manner, the external overcomplete dictionary is compressed from 360 elements, to have the same number of atom in each experiment.

Table 1 shows the accuracy of ESRC and $RADL_{wo}$ classification with different size of external data and compressed data. The average accuracy is shown for 10 experiments. Because the amount standard deviation is negligible, it is not reported in the Table 1. We notice that by increase the number of atoms, the performance increases monotonically in general. Moreover the overall performance is higher for GCDL than methods without external dictionary learning. The result shows that in low compression rate (equivalent to high number of atoms), accuracy is low compare to other methods. Since GCDL is based on cluster-

---

[1] http://mml.citi.sinica.edu.tw/publications.html

ing, the amount of information added to enhance SRC declines as the # of atoms grows. In another words, by increasing the number of atoms close to the total number of data clusters, we tend to partition data randomly without using structural information of data.

In the second experiment, the performance of GCDL is compared to some dictionary learning methods followed by classification algorithms such as $RADL_wDL$ and SVDL [2] , then with K-SVD [3] , MOD and SMRS are used for dictionary learning followed by $RADL_{wo}$ as classification. Results of ESRC and $RADL_{wo}$ are provided by applying whole external data without dictionary learning. Finally, results are provided for SRC and RSC as a base line. Since these methods do not require external dictionary. The execution time of these methods are also compared. The codes are implemented in MATLAB , using a 3.40 GHz and 8 GB RAM computer.

Table 2 shows that the average execution time of constructing compressed dictionary by GCDL is much lower than other methods. In addition, our method doesn't rely on parameter tuning like K-SVD, MOD and SMRS. Finally, GCDL provides the best accuracy.

6 CONCLUSIONS

In this paper, the Graph Compression Dictionary Learning (GCDL) method based on matrix factorization is proposed to construct compact representation of overcomplete external data. The GCDL exploits the structural information on external dictionary to build compressed dictionary. It affects a trade-off between time complexity and accuracy. Experiment conducted with a high compression rate produces a better accuracy. Therefore we show that it is more robust to intra class variation compare into common used dictionary learning methods on literature. As a result, the proposed algorithm allows managing occlusion face images, or illumination and expression. Moreover, GCDL handles even one or few gallery images per individuals. The result on CAS-PEAL dataset show that GCDL has a better time efficiency for the construction of compact dictionary. It can be employed to accelerate many SRC approaches, and the complexity of a matrix-vector multiplication can be significantly reduced.

---

[2]http://www4.comp.polyu.edu.hk/ cslzhang/papers.htm
[3]http://www.cs.technion.ac.il/ elad/software/

REFERENCES

Aharon, M., Elad, M., and Bruckstein, A. (2006). The k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Trans. Signal Processing*, 54(11):4311–4322.

Bashbaghi, S., Granger, E., Sabourin, R., and Bilodeau, G. (2014). Watch-list screening using ensembles based on multiple face representations. In *International Conference on Pattern Recognition*, pages 4489–4494.

Choi, Y. and Szpankowski, W. (2012). Compression of graphical structures: Fundamental limits, algorithms, and experiments. *IEEE Trans. on Information Theory*, 58(2):620–638.

Deng, W., Hu, J., and Guo, J. (2012). Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans. Pattern Analysis Machine Intelligence*, 34(9):1864–1870.

Dewan, M. A., Granger, E., Marcialis, G. L., Sabourin, R., and Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition*, 49:129–151.

Donoho, D. L. and Tsaig, Y. (2008). Fast Solution of l1-Norm Minimization Problems When the Solution May Be Sparse. *Information Theory, IEEE Transactions on*, 54(11):4789–4812.

Elhamifar, E., Sapiro, G., and Vidal, R. (2012). See all by looking at a few: Sparse modeling for finding representative objects. In *IEEE Conference on Computer Vision and Pattern Recognition,*, pages 1600–1607.

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., and Zhao, D. (2008). The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Trans. System Man Cybernetics Part A*, 38(1):149–161.

Lee, D. D. and Seung, H. S. (2001). Algorithms for nonnegative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562.

Mairal, J., Bach, F., and Ponce, J. (2014). Sparse modeling for image and vision processing. *Foundations Trends in Computer Graphics and Vision*, 8(2-3):85–283.

Mokhayeri, F., Granger, E., and Bilodeau, G. (2015). Synthetic face generation under various operational conditions in video surveillance. In *International Conference on Image Processing.*

Navlakha, S., Rastogi, R., and Shrivastava, N. (2008). Graph summarization with bounded error. In *International Conference on Management of Data (ACM)*, pages 419–432.

Nourbakhsh, F. (2015). *Algorithms for Graph Compression: Theory and Experiments*. PhD thesis, Dipartamento di Scienze Ambientali, Informatica e Statistica, Universitá Ca'Foscari, Venice, Italy.

Nourbakhsh, F., Bulò, S. R., and Pelillo, M. (2015). A matrix factorization approach to graph compression with partial information. *International Journal of Machine Learning & Cybernetics*, 6(4):523–536.

Ramírez, I., Sprechmann, P., and Sapiro, G. (2010). Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE*

Methods	Number of Atoms					
with External Dictionary	6	12	18	30	60	90
ESRC	74.50%	74.50%	75.17%	75.83%	77.33%	78.33%
ESRC+**GCDL**	74.90%	75.57%	76.08%	76.35%	77.62%	77.85%
$RDAL_{wo}$	83.17 %	83.83%	84.33 %	84.00 %	85.90 %	86.83 %
$RDAL_{wo}$+**GCDL**	83.95%	84.87%	85.53%	85.15%	85.97%	86.73%

Table 1: Average classification accuracy of GCDL and references methods on CAS-PEAL data.

Methods with Dictionary Learning and Classification				
	Number of Atoms	Size of Dictionary	Accuracy	Time in Sec
SVDL	19	2000×360	82.33 %	1607.87
$RADL_wDL$	18	2000×360	85.67%	924.62
Methods with Dictionary Learning and $RADL_{wo}$ as Classification				
K-SVD	18	2000×360	84.96%	115.38
MOD	18	2000×360	85.17%	71.759
SMRS	17	2000×360	83.83%	3.28
Methods with Compression and $RADL_{wo}$ as Classification				
GCDL	18	2000×360	**86.00%**	**0.87**
Apply whole Raw Dictionary				
ESRC	360	2000×360	78.50%	169.69
$RADL_{wo}$	360	2000×360	87.00%	267.53
Methods without External Dictionary				
SRC	72.67%			
RSC	82.33%			

Table 2: Average accuracy and reconstruction time of GCDL and references methods on CAS-PEAL data.

Conference on Computer Vision and Pattern Recognition, pages 3501–3508.

Shafiee, S., Kamangar, F., Athitsos, V., and Huang, J. (2013). The role of dictionary learning on sparse representation-based classification. In *International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, pages 47:1–47:8.

Sperotto, A. and Pelillo, M. (2007). Szemerédis regularity lemma and its applications to pairwise clustering and segmentation. In *International Conference of Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 4679 of *Lecture Notes in Computer Science*, pages 13–27.

Szemerédi, E. (1978). Regular partitions of graphs. In *Problèmes combinatoires et thorie des graphes*, pages 399–401.

Tan, X., Chen, S., hua Zhou, Z., and Zhang, F. (2006). Face recognition from a single image per person: A survey. *International Journal of Pattern Recognition*, 39:1725–1745.

Tillmann, A. M. (2015). On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letter*, 22(1):45–49.

Toivonen, H., Zhou, F., Hartikainen, A., and Hinkka, A. (2011). Compression of weighted graphs. In *International Conference on Knowledge Discovery and Data Mining (ACM)*, pages 965–973.

Wei, C. and Wang, Y. F. (2015). Undersampled face recognition via robust auxiliary dictionary learning. *IEEE Transactions on Image Processing*, 24(6):1722–1734.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis Machine Intelligence*, 31(2):210–227.

Yang, A. Y., Ganesh, A., Zhou, Z., Sastry, S., and Ma, Y. (2010a). Fast l_1-minimization algorithms for robust face recognition: A review. *International Conference on Image Processing*, pages 1849–1852.

Yang, M., Van, L., and Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. In *International Conference on Computer Vision*, pages 689–696.

Yang, M., Zhang, L., Feng, X., and Zhang, D. (2011a). Fisher discrimination dictionary learning for sparse representation. In *International Conference on Computer Vision*, pages 543–550.

Yang, M., Zhang, L., Yang, J., and Zhang, D. (2010b). Metaface learning for sparse representation based face recognition. In *International Conference on Image Processing,*, pages 1601–1604.

Yang, M., Zhang, L., Yang, J., and Zhang, D. (2011b). Robust sparse coding for face recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 625–632.