A JOINTLY LOCAL STRUCTURED SPARSE DEEP LEARNING NETWORK FOR FACE RECOGNITION

Renjie Wu, Sei-ichiro Kamata

Graduate School of Information, Production and Systems, Waseda University, Japan

ABSTRACT

In this paper, we proposed an optimized Sparse Deep Learning Network (SDLN) model for Face Recognition (FR). A key contribution of this work is to learn feature coding of human face with a SDLN based on local structured Sparse Representation (SR). In traditional sparse FR methods, different poses and expressions of training samples could have great influence on the recognition results. We consider the SR that should be guided by context constraints which are defined by the correlations of dictionary atoms. The over-complete common dictionary that contains common atom set has been learned from a local region structured sparse encoding process. We obtained over-complete common dictionary and feature coding for each face. As we all know that the deep learning has been widely applied to face feature learning. Using traditional deep learning methods can not contain variations of face identity information. We have to get face features of compatible change in a jointly deep learning network. The proposed SDLN is jointly fine-tuned to optimize for the task of FR. The SDLN achieves high FR performance on the ORL and FERET database.

Index Terms— Face recognition, Atom decomposition, Over-complete dictionary learning, Sparse deep learning network, Restricted boltzmann machine

1. INTRODUCTION

Face Recognition (FR) has become one of the hottest topics in the area of image processing. It has been extensively applied in identity validation and recognition. Many researchers to study new FR algorithms [1, 2, 3, 4, 5, 6] for years. The most arresting features of FR is its non-intrusive and non-contact property. Human face has advantage over other biometric information because it is natural and easy to use biometric recognition. In this field, the key challenge of FR is to get identity information of object face in large different poses, illuminations, expressions, ages, makeups, and occlusions.

In traditional FR methods, generally, we have two problems of the recognition process. First is human face feature extraction problem. In first step, many feature extraction methods are proposed to mark the feature points, but their methods cannot extract the directly relation of human face identity. This means some important information have been lost at the feature extraction stage. For example, Scale Invariant Feature Transform (SIFT) [7] was used for extracting highly discriminative local image features. These features are invariant to image scaling and rotation, and partially invariant to changes in illumination and viewpoint. These local image features do not have the abstract concept of human face. Other some learning-based feature extraction approaches are proposed. Their optimization functions are defined in the sub-block of face image [8, 9].

The next problem is face classification (training stage) and recognition (testing stage). In this stage, the features of the unknown sample are first described with the same features extraction methods in the training stage. And the features coding of the testing sample are compared with training samples, and then the sample is assigned to one category. If the features of the testing sample are matching to be optimum, then it can be identified. There, if training data have labeled data, then we can call 'supervised'. Unlabeled training data learning method such as Support Vector Machines (SVM) [8, 9, 10] called 'unsupervised' learning. Of course, for other discriminative methods, we can calculate the similarities of some faces to recognize testing objects [5, 11]. However, these tradition methods do not have a deep network. This means some important information will be lost in the learning process. And, the feature extraction and recognition step cannot be jointly optimized, because they are independent process.

So, all of the issues discussed above need to learn a jointly learning network to handle FR. In this paper, we use SR to perform feature extraction. We also can represent testing samples in over-complete common dictionary and sparse features. If training samples are sufficient, then this representation is naturally sparse. In usual sense, we need use low-dimensional features of face to express most relevant informative in FR. Tradition methods have been devoted to investigate feature transformations for projecting the high-dimensional into lower dimensional feature spaces (e.g., Eigenfaces [12], Fisherfaces [13], Laplacianfaces [14], and variants [15, 16]). For Sparse Representation (SR), extracted features contain enough information to recover feature face and correctly classify any object face. And the sparse features also contain the identity information of human faces.

This work was supported by JSPS KAKENHI Grant Number 15K00248.

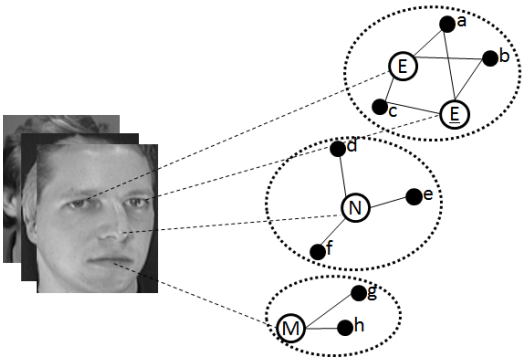

Fig. 1. Graph Structure of face. The edges describe that there exist some context correlations between local features and dictionary subgraphs.

Of course, the dimension of the feature space should be sufficiently large.

The SR should be guided by the context constraints for more smooth expression of face information. The context constraints can be defined by the correlations of feature points. In this case, we used a structured model to describing the sparse constraints. We propose a Graph Structure of face as show in Fig. 1. The nodes $\{E, \underline{E}, ..., M\}$ are feature nodes corresponding to the correlated outputs (nodes $(a, b, ..., h)$). E nodes (feature nodes of left eye) are much more similar with $\underline{E}$ nodes (feature nodes of right eye). Therefore, the same set of dictionary atoms and similar coefficients will be learned in a neighbor feature nodes space. Finally, the Sparse Deep Learning Network (SDLN) model must be correctly computed. The process through multiple layers of the deep belief networks (DBNs) [17] is to extract high-level (abstractly) features. In the experiment, we use softmax regression method to training the classifier. Many Restricted Boltzmann Machine (RBM) layers of DBNs can extract different levels of features. For our proposed optimized SDLN, the feature extraction and classification are jointly optimized by learning network. Our SDLN as show in Fig. 2. The experimental results demonstrate the effectiveness of the proposed algorithm.

2. RELATED WORK

2.1. Sparse Representation [18]

Firstly, we need to define the over-complete dictionary $D \in R^{J \times N}$ containing the atoms, and an object face $y \in R^J$. In this paper, J is the total number of features of the face samples. The problem of the SR is to find a coefficient vector $s \in R^N$ such that $y = Ds$ and $\|s\|_1$ is minimized:

$$\min_{s,\lambda} \frac{1}{2} \|y - Ds\|_2^2 + \lambda \|s\|_1 , \quad (1)$$

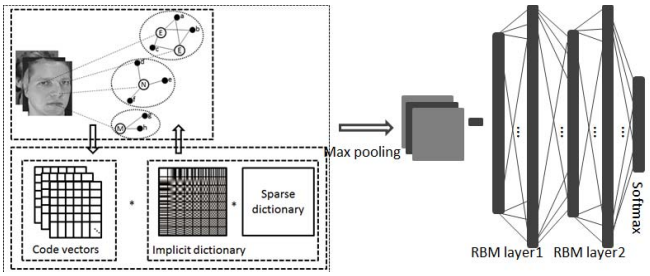

Fig. 2. Jointly local structured SDLN. The arrows show forward and back propagation directions.

where $\|*\|_1$ is the l_1-norm. The parameter λ is a scalar regularization parameter that balances the trade-off between reconstruction error and sparsity. We can use the Bayesian method [19] to learning the optimal value of λ.

2.2. Restricted Boltzmann Machine [17]

We extract relational visual features (visible layer) from multiple faces. The high-layer RBM in our framework is similar to the DBNs. The RBM is a two-layer, bipartite, undirected graphical model. It contains a set of binary hidden units h, a set of (binary or real-valued) visible units v, and symmetric connections between these two layers represented by a weight matrix W. The probability distributions are defined as follows:

$$P(v, h) = \frac{e^{-E(v,h)}}{Z}, \quad (2)$$

where Z is the partition function. If the visible units are binary-valued, we define the energy function of a configuration as:

$$E(v, h) = -\Sigma_{i,j} v_i W_{ij} h_j - \Sigma_j b_j h_j - \Sigma_i c_i v_i, \quad (3)$$

where b_j are hidden unit biases and c_i are visible unit biases. Form the energy function, we can easy to know that the hidden units are conditionally independent of one another given the visible layer, and vice versa.

3. JOINTLY LOCAL STRUCTURED SDLN

3.1. Local Structured Sparse Representation of Face

SR has been widely used in the FR to learn the most discriminant visual features and face identity classification. The optimization problem in Eq. (1) can be solved by the lasso algorithm [21]. If it is regularized by the l_1-norm, then sparsity coefficients are estimated by each training face region in D individually. The relationships and structures of face has neglected in Eq. (1).

For lasso algorithm, we can see that the group lasso penalty can be calculate as imposing l_2-norm on the group

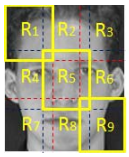

Fig. 3. human faces into 9 overlapping regions.

level. In this paper, we divide human faces into 9 overlapping regions as Fig. 3. Then the same feature blocks in different human faces could be jointly learned by SR. We need to define the similar feature blocks of the same location. Firstly, the features within the same face could be different in visual. Secondly, the feature blocks can be translated and scaled by feature center. Finally, the features within the same face could be selected by different groups.

Given a face image dataset $D \in R^{J\times N}$ with N images. Then we can formulate a SR framework as follows:

$$\min_{s,\lambda} \frac{1}{2} \left\| y - \Sigma_{n=1}^{N} D_n s_n \right\|_2^2 + \lambda \Sigma_{r=1}^{9} \Sigma_{n=1}^{N} \left\| s_{rn} \right\|_2 , \quad (4)$$

where $\lambda \geq 0$ is the regularization parameter for the group lasso penalty. And all blocks in D can be partitioned into 9 overlapping groups of N non-overlapping groups.

3.2. Internal Structure of Local Region

Now we need to consider the problem of feature nodes structure in the local region. For the feature nodes in the overcomplete dictionary D, let us define the structure of the F_r features is available in a region R with a set of feature nodes $V_r = \{1, 2, ..., F_r\}$ and a set of edges E_r. Let $\omega_{a_r b_r} \geq 0$ denote the weight of the edge $e_r = (a_r, b_r) \in E_r$, corresponding to the correlation between two nodes a and b. Here, $\omega_{a_r b_r} \geq 0$ (positively correlated image feature).

In order to integrate the internal structure into our local sparse and guide sparsely feature extraction, we can rewrite the Eq. (4), and propose the local internal structure SR as follows:

$$\begin{aligned} &\min_{s,\gamma,\lambda} \frac{1}{2} \left\| y - Ds \right\|_2^2 + \lambda \left\| s \right\|_1 \\ &+ \gamma \Sigma_{r=1}^{9} \Sigma_{e_r=(a_r,b_r)\in E, a_r<b_r} \omega_{a_r b_r} |s_{a_r} - s_{b_r}|, \end{aligned} \quad (5)$$

where s_{a_r} and s_{b_r} are coefficients in s_r corresponding to the selection coefficients of the a_rth and b_rth feature nodes in region r, γ is the regularization parameter. Therefore, we can get a densely connected subnetwork of highly correlated features in region R. Some similarity features can be jointly selected into same group.

3.3. External Structure between Different Nearby Local Regions

We also need to consider the fusion of similar feature nodes between different nearby local regions. We can get a better performance of sparse feature extraction, if we construct the appropriately nearby Interregional structure. Then, we learned some high-level relationship features of eyes, nose, mouth and ears in different regions into the same structure group.

Introducing nearby interregional structure, we can rewrite the Eq. (5) as follows:

$$\begin{aligned} &\min_{s,\gamma,\kappa,\lambda} \frac{1}{2} \left\| y - Ds \right\|_2^2 + \lambda \left\| s \right\|_1 \\ &+ \gamma \Sigma_{r=1}^{9} \Sigma_{e_r=(a_r,b_r)\in E, a_r<b_r} \omega_{a_r b_r} |s_{a_r} - s_{b_r}| \\ &+ \kappa \Sigma_{\hat{r}=1}^{12} \Sigma_{e_{\hat{r}}=(a_{\hat{r}},b_{\hat{r}})\in E_{\hat{r}}, a_{\hat{r}}<b_{\hat{r}}} \omega_{a_{\hat{r}} b_{\hat{r}}} |s_{a_{\hat{r}}} - s_{b_{\hat{r}}}|, \end{aligned} \quad (6)$$

where $\hat{r}$ stand for 12 nearby regions, κ is the regularization parameter that control the complexity of the structure model.

The optimization problem in Eq. (6) is convex. For the optimization function of Eq. (6), the main problem is non-separability of s in the non-smooth penalty function ω. In this section, we use a general smoothing proximal gradient (SPG) [22] method to solve Eq. (6).

The SPG first make a separable and smooth approximation of ω, and then solves this transformed simple l_1-norm penalized sparse learning problem by the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [23]. Then we need to solve the following optimization problem:

$$\begin{aligned} &\min_{s,\gamma,\kappa,\lambda} \frac{1}{2} \left\| y - Ds \right\|_2^2 + \gamma \Sigma_{r=1}^{9} f_\mu(s_r) \\ &+ \kappa \Sigma_{\hat{r}=1}^{12} f_\mu(s_{\hat{r}}) + \lambda \left\| s \right\|_1 , \end{aligned} \quad (7)$$

where the approximate smoothing approximation function $f_\mu(s_r)$ as follows:

$$f_\mu(s_r) = max_{\alpha\in\{\alpha | \|\alpha\|_\infty \leq 1, \alpha\in R^{|E_r|}\}} \alpha Q s_r - \mu 0.5 \|\alpha\|_2^2, \quad (8)$$

where $\alpha \in \{\alpha | \|\alpha\|_\infty \leq 1, \alpha \in R^{|E_r|}\}$ is a vector of auxiliary variables associated with $\|Qs\|_1$, $Q \in R^{|E_r|\times F_r}$ is the edge-vertex incident matrix, $\mu \geq 0$ is the smoothness parameter.

The first two terms of Eq. (7) are convex and smooth, then, Eq. (7) can be efficiently solved by the FISTA algorithm. So, the SPG can achieve $O(1/\epsilon)$ convergence rate for a desired accuracy ϵ.

3.4. Jointly Learning Network

In this paper, we use the SIFT and $k-$nearest neighbor ($k-$NN) to calculating the relationship between the nearby regions. There are some traditional methods that used the deep learning models for FR, but the extracted features are independent from each face. The relations between the two

Fig. 4. Samples from ORL database.

faces are not included at their feature extraction steps. We directly extract visual features from the jointly learning network. The high-layer RBM in our network is similar to that of the DBNs.

4. EXPERIMENTS

In this paper, we perform FR on benchmark face databases, including ORL and FERET to demonstrate the performance. We first discuss the parameter setting. And then, we test the different poses of human face on ORL database and show the robustness. Final, we conduct experiments on FERET database. In this paper, we used the method to initialize the DCT dictionary to better learn the dictionary as follows

$$D_r = A * \tilde{D}_r \tag{9}$$

where A is DCT dictionary, $\tilde{D}_r$ is sparse dictionary.

4.1. Parameter Setting

In sparse layer, the regularization parameters $\gamma = 1e-5; \kappa = 1e-6; \lambda = 1e-3$ are chosen by cross-validation to fix the sparsity of all the SR methods in this experiment.

4.2. ORL Database

We first test the performance of all the competing methods by using the ORL database. The ORL database consists of facial images of 40 different people and each individual has 10 different images. All images are gray scale and normalized to a resolution of 112×92 pixels. The samples as Fig. 4

In this experiment, 5 images of each person will be randomly chosen for training, while the remaining images for testing. To compare our method and other state-of-the-art approaches, 5 tests are performed with a 5 training samples. From Table. 1, we see that our recognition rates are higher than all the other competing methods. We may find that the methods with learning method are significantly better than the ones without learning.

4.3. FERET Database

We then conduct FR on the FERET database. The FERET images contain variations in lighting, facial expressions, pose angle, etc. In this experiment, each face is resized to 100×100 pixels. Fig. 5 show the samples. We use set 'Fa' for training, and 'Fb', 'Dup I' and 'Dup II' for testing.

Table 1. FR rates (complete face test) on ORL database.

Method	rates	Method	rates
NN	0.960	ESRC [24]	0.980
SVM	0.970	ESRC-KSVD [25]	0.980
SRC [26]	0.970	SVDL [25]	0.980
AGL [27]	0.975	Our Method	0.985

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

Fig. 5. Sample of FERET database, (a) 'Fa' set, (b) 'Fb' set, (c) 'Dup I' set, (d) 'Dup II' set.

In FERFT experiment, we use images of 'Fa' set to training data, and test the 'Fb', 'Dup I', 'Dup II' set. For this experiment, we compare to some start-of-the-art methods. From Table. 2, we can see that the our method could achieve higher recognition rates than the other methods with face features on FERET database. This validates that the jointly local structured SDLN is more better than these tradition methods.

Table 2. FR rates on FERET database from cropped images.

Method	Fb	Dup I	Dup II
NN	0.868	0.466	0.301
SVM	0.873	0.467	0.303
SRC [26]	0.872	0.486	0.335
AGL [27]	0.947	0.579	0.385
ESRC [24]	0.912	0.602	0.518
ESRC-KSVD [25]	0.956	0.602	0.557
SVDL [25]	0.958	0.652	0.592
Our Method	0.969	0.679	0.657

5. CONCLUSION

We proposed a jointly local structured SDLN. This method jointly extracts local structured correlated visual features to improve FR performance. Both feature extraction and recognition are jointly optimized under the SDLN. The experiments with FR validated the performance of proposed network on ORL and FERET database. However, a problem is that if the training sample is not sufficient, it cause learned dictionary is not over-complete. We plan to use low-rank recovery to solve this problem.

6. REFERENCES

[1] X. Wang and X. Tang, "Random sampling for subspace face recognition," *IJCV*, vol. 70, pp. 91 – 104, May. 2006.

[2] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *PAMI*, vol. 28, pp. 2037 – 2041, Dec. 2006.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, , and S. K. Nayar, "Attribute and simile classifiers for face verification," *ICCV*, pp. 365 – 372, Oct. 2009.

[4] Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," *CVPR*, pp. 497 – 504, Jun. 2013.

[5] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," *CVPR*, pp. 3025 – 3032, Jun. 2013.

[6] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," *ICCV*, pp. 113 – 120, Dec. 2013.

[7] L. Zhang, J. Chen, Y. Lu, and P. Wang, "Face recognition using scale invariant feature transform and support vector machine," *ICYCS*, pp. 1766 – 1770, Nov. 2008.

[8] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," *CVPR*, pp. 2707 – 2714, Jun. 2010.

[9] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," *CVPR*, pp. 2518 – 2525, Jun. 2012.

[10] N. Pinto and D. D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," *FG*, pp. 8 – 15, Mar. 2011.

[11] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," *ACCV*, pp. 709 – 720, Nov. 2010.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *CVPR*, pp. 71 – 86, Jan. 1991.

[13] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *PAMI*, vol. 19, pp. 711 – 720, Jul. 1997.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *PAMI*, vol. 27, pp. 328 – 340, Mar. 2005.

[15] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ica for face recognition robust to local distortion and partial occlusion," *PAMI*, vol. 27, pp. 1977 – 1981, Mar. 2005.

[16] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," *CVPR*, vol. 1, pp. 207 – 212, Dec. 2001.

[17] G. Hinton and S. Osindero, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527 – 1554, Jul. 2006.

[18] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, pp. 2845 – 2862, Nov. 2001.

[19] Y. Lin and D. Lee, "Bayesian l1-norm sparse learning," *ICASSP*, vol. 5, pp. V – V, May. 2006.

[20] Xi Chen, Seyoung Kim, Qihang Lin, Jaime G. Carbonell, and Eric P. Xing, "Graph-structured multi-task regression and an efficient optimization method for general fused lasso," *CoRR*, May. 2010.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B(Methodological)*, vol. 58, pp. 267 – 288, 1996.

[22] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing, "An efficient proximal gradient method for general structured sparse learning," *stat 1050*, vol. 26, 2011.

[23] A. Beck and M. Teboulle, "A fast iterative shrinkagethresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183 – 202, Mar. 2009.

[24] W. H. Deng, J. N. Hu, and J. Guo, "Extended src: Undersampled face recognition via intra-class variant dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1864 – 1870, Jan. 2012.

[25] M Yang, Luc Van Gool, and L Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," *ICCV*, pp. 689 – 696, Dec. 2013.

[26] J. Wright, A. Y. Yang, S. S. Sastry, and Y. Ma, "Robust face recogntion via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210 – 227, Apr. 2009.

[27] Y. Su, S. Shan, X. Chen, and W. Gao, "Adaptive generic learning for face recognition from a single sample per person," *CVPR*, pp. 2699 – 2706, Jun. 2010.