LETTER

Learning Convolutional Domain-Robust Representations for Cross-View Face Recognition

Xue CHEN[†a)], **Chunheng WANG**[†], **Baihua XIAO**[†], ***Nonmembers***, ***and*** **Song GAO**[†], ***Member***

SUMMARY This paper proposes to obtain high-level, domain-robust representations for cross-view face recognition. Specially, we introduce Convolutional Deep Belief Networks (CDBN) as the feature learning model, and an CDBN based interpolating path between the source and target views is built to model the correlation of cross-view data. The promising results outperform other state-of-the-art methods.

key words: *cross-view, face recognition, convolutional deep belief networks, domain-robust*

1. Introduction

Automatic face recognition systems can achieve high performance under frontal view. However, in real scenarios, face images are generally captured under various views, which degrades the performance severely. The difficulty for cross-view face recognition is that the view varies in 3D space, while the image captures only 2D appearances. As the view changes, different visible parts of face appear in the images. This leads to a special phenomenon that faces of different identities with similar views are more similar than that of the same identity under different views. The difference brought by variant views could be larger than that caused by identity changes, making cross-view face recognition problem very difficult.

To address this problem, one popular family of statistic-based learning methods aim at seeking view-specific transforms and then project the samples into a common subspace. Typically, Lin *et al.* [1] proposed Common Discriminant Feature Extraction (CDEF) to transform samples of different modalities to the common feature space. Sharma *et al.* [2] and Li *et al.* [3] introduced Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) to maximize the intra-individual correlation of varying-view faces in the mapping space. Moreover, extensions of such pairwise methods are also developed for multi-view problems, such as Multi-view Discriminant Analysis (MvDA) [4], and Multi-view CCA [5].

However, there are some limitations for methods above. One limitation is that they pursue linear transforms to construct the projection space, which often severely limits the capacity of representations. The other limitation is that they learn the view-specific transform based on single-view data, neglecting the correlation information of cross-view data. For these limitations, many approaches have been developed. Typically, deep networks have achieved tremendous success on many learning tasks for the ability of learning powerful non-linear representations. Deep Belief Networks [6] and Convolutional Deep Belief Networks [7] are two popular models to construct such deep learning architectures. On the other hand, to draw the statistical connections between cross-view data, the domain adaptation methods have shown promising results by exploring a virtual interpolating path between views. For example, Li *et al.* [8] modeled each virtual view as a linear transformation of the descriptor, and representations built upon the virtual path aimed at bridging the source and target views. Chopra *et al.* [9] proposed a deep-model based interpolating path to learn predictive representations by exploiting the distribution shift information between domains (hereafter referred to as DLID). Specially, the basic deep sub-model in DLID is composed of four components cascaded together, namely Filtering (F), Rectification (R), Normalization (N), Pooling (P), abbreviated as F-R-N-P [10].

In this paper, we propose a novel feature learning method to obtain high-level, domain-robust representations for cross-view face recognition. It combines ideas from both of the previous approaches. The architecture of our model is illustrated in Fig. 1. First, we introduce Convolutional Deep Belief Networks (CDBN) as the transform model to learn

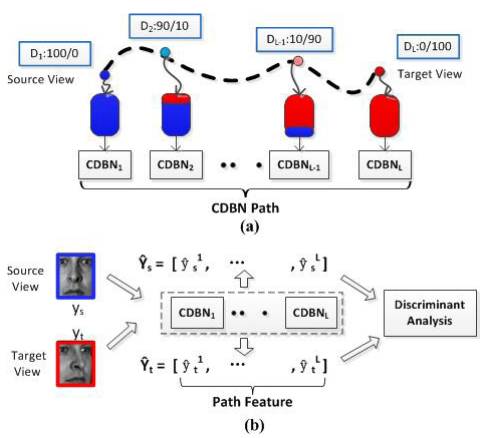

Fig. 1 (a) The interpolating path between the source and target views. Each intermediate dataset $\{D_l : a/b\}$ is created by sampling $a\%$ of the source data and $b\%$ of the target data. An CDBN is trained on each dataset D_l. (b) For cross-view data $\{y_s, y_t\}$, $\{\hat{y}_s^l, \hat{y}_t^l\}$ is the non-linear feature from model $CDBN_l$. We apply discriminant analysis on the path feature $\{\hat{Y}_s, \hat{Y}_t\}$.

Manuscript received May 13, 2014.
Manuscript revised August 8, 2014.
Manuscript publicized September 8, 2014.

[†]The authors are with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
a) E-mail: xue.chen@ia.ac.cn

DOI: 10.1587/transinf.2014EDL8095

hierarchical non-linear representations of the inputs. Specially, CDBN scales to realistic image sizes and captures the spatial correlation of images effectively. Second, to model the correlation of cross-view data, we define an interpolating path by sampling intermediate datasets along the distribution shift between the source and target data. An CDBN is trained on each intermediate dataset. All the outputs from the CDBNs are concatenated as a path feature for the input, which is highly rich to model the source to target transition information. Finally, the discriminant analysis is applied on the resulting features in order to gain more discriminative power and be suitable for cross-view recognition task.

The rest of the paper is organized as follows. Section 2 gives the details of the proposed method. Section 3 demonstrates that our experimental results are more accurate than state-of-art methods on CMU Multi-PIE dataset. Finally, in Sect. 4 we conclude the paper.

2. Methods

2.1 Convolutional Deep Belief Networks

For computer vision problems, convolutional networks seem a natural choice to capture high-level hierarchical representations of images. In this paper, we adopt Convolutional Restricted Boltzmann Machine (CRBM) [7] as the basic sub-model, and construct CDBN as the deep convolutional model by stacking CRBMs hierarchically. Specially, CRBM is a probabilistic generative model which is optimized to maximize the likelihood of the training data. From the probabilistic modeling perspective, feature learning based on CRBM is to recover a set of latent hidden variables that describe the distribution of the observed data. Based on this, applying CRBM for face feature learning can effectively explore the latent explanatory factors of variations on face images (eg, pose variations and identity variations) and further model the characteristics of face data accurately [11], [12]. Following, we describe the CRBM model and the construction of CDBN by CRBMs in detail.

The basic CRBM consists of two layers: a visible layer V and a hidden layer H. The visible layer is an $N_V \times N_V$ array of binary units. The hidden layer consists of K groups of $N_H \times N_H$ arrays of binary units. Each of the K groups is associated with a $N_W \times N_W$ filter ($N_W = N_V - N_H + 1$); the filter weights W^k are shared across all the hidden units within the group. In addition, each hidden group has a bias b_k and all visible units share a single bias c. To make the CRBM more scalable and incorporate local translation invariance, a probabilistic max-pooling layer P is generally added as the last layer, where a $C \times C$ block of hidden units are shrunk to a pooling node by computing the maximum [7], [12]. An illustration of CRBM is shown in Fig. 2.

To deal with real-value data, our CRBM model uses gaussian units for visible variables and binary units for hidden variables. The energy function is defined as:

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h)) \tag{1}$$

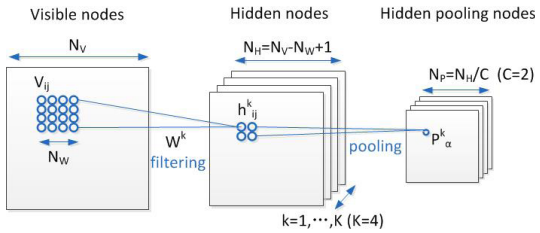

Fig. 2 Schematic diagram of CRBM. The visible unites $v_{i,j}$ in layer V are convolved with the filter W^k to be the hidden units $h^k_{i,j}$ in layer H. Further, the $C \times C$ block of hidden units $h^k_{i,j}$ in layer H are shrunk to the pooling nodes p^k_α in layer P. For illustration, we set $K = 4$ and $C = 2$.

$$\begin{aligned} E(v,h) = &-\sum_{k=1}^{K}\sum_{i,j=1}^{N_H}\sum_{r,s=1}^{N_W} h^k_{i,j} W^k_{r,s} v_{i+r-1,j+s-1} \\ &+ \sum_{i,j=1}^{N_V}\frac{1}{2}v^2_{i,j} - \sum_{k=1}^{K} b_k \sum_{i,j=1}^{N_H} h^k_{i,j} - c\sum_{i,j=1}^{N_V} v_{i,j}, \\ s.t. &\sum_{(i,j)\in B_\alpha} h^k_{i,j} \le 1, \forall k, \alpha \end{aligned} \tag{2}$$

where B_α refers to a $C \times C$ pooling block of hidden units $h^k_{i,j}$ connecting a pooling node p^k_α. Under the energy function, the conditional distributions $P(v_{i,j} = 1|h)$ and $P(h^k_{i,j} = 1|v)$ are easy to compute. The CRBM can be trained like the standard RBM using the contrastive divergence (CD) algorithm [7], [13]. After training a CRBM, the hidden (pooling) activations are used as input to further train a next layer CRBM. Finally, an CDBN is constructed by stacking these pre-trained CRBMs hierarchically. Specially, by setting the visible layer V of the energy function in Eq. (2) as realistic image sizes, we can train CRBMs and further construct CDBN on full-sized images easily.

Typically, by performing hierarchical (bottom-up and top-down) inference over full-sized face images, CRBM learns semantically meaningful visual features such as edges, face parts at different hierarchies by exploiting the implicit structure of face images [7]. A diagram of the hierarchical feature architecture in CDBN is shown in Fig. 3. These low-level to high-level visual features are combined to describe the face characteristics in a hierarchical and complementary way. In addition, the spatial correlation of different face parts (eg, the relative positions of eye and nose, or nose and mouth) is very important for describing the fine face characteristics. As shown in the red boxes of Fig. 3, some high-level features learned from CDBN characterize neighbouring parts of faces (eg, eye and nose, or nose and mouth), and hence capture the spatial information of these face parts potentially. From these points, CDBN is just appropriate for modeling the characteristics of face images.

2.2 CDBN Based Interpolating Path

Exploring a potential transition path between the source and target data is a popular way to model the correlation information of cross-domain data. In this paper, we introduce a different notion of interpolating path, by sampling interme-

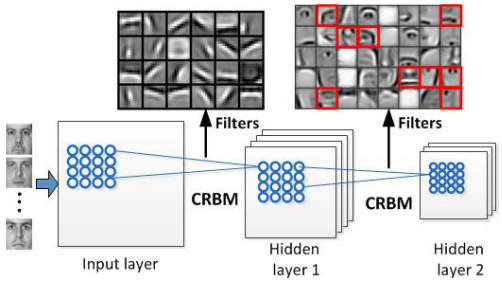

Fig. 3 The hierarchical face feature architecture in CDBN. Each sub-image on the top is the visualization of the filter weight learned. The low-hierarchy (Layer 1) CRBM learns edge-like features; the high-hierarchy (Layer 2) CRBM learns part-like visual features. Specially, the filters in the red boxes characterize neighbouring face parts (eg, eye and nose, or nose and mouth).

diate datasets along the distribution shift between the source and target data. Moreover, we train an CDBN for each dataset on the path. The resulting representation based on the CDBN path is highly rich in containing the source to target path information and robust to the variations caused by domain changes.

Assume the dataset of the source domain S as D_S, and the dataset of the target domain T as D_T. Starting with D_S, we generate intermediate datasets, where for each successive dataset we gradually increase the proportion of samples randomly drawn from D_T, and decrease the samples drawn from D_S, as shown in Fig. 1 (a). In particular, let $l \in [1, \ldots, L]$ be an index over the L datasets we generate. Then we have $D_1 = D_S$, $D_L = D_T$. For $l \in [2, \ldots, L-1]$, datasets D_l and D_{l+1} are created in a way so that the proportion of samples from D_T in D_l is less than in D_{l+1}. Each of these datasets can be thought of as a single point on a particular kind of interpolating path between S and T.

Next, we train an CDBN for each intermediate dataset D_l on the path in an unsupervised way, as described in Sect. 2.1. In this way, an optimized CDBN path $[CDBN_1, CDBN_2, \ldots, CDBN_L]$ is built to connect the source and target domains. Actually, each basic $CDBN_l$ model on the path can be considered as a deep nonlinear feature extractor F_{W_l}. For an one-layer CDBN model, a fast bottom-up inference of the hidden layer H in group k conditioned on the visible layer V is computed as:

$$I(h^k_{i,j}) = b_k + (\tilde{W}^k *_v v)_{i,j}, \tag{3}$$

$$P(h^k_{i,j} = 1|v) = \frac{\exp(I(h^k_{i,j}))}{1 + \sum_{(i',j') \in B_\alpha} \exp(I(h^k_{i',j'}))}, \tag{4}$$

where $\tilde{W}^k$ is the 2-d filter matrix W^k flipped vertically and horizontally, and $*_v$ denotes valid convolution. The sub-sampled activation probabilities $P(h^k_{i,j} = 1|v)$ act as the feature for layer H. In this setup, for a M-layer $CDBN_l$, the nonlinear feature extractor F_{W_l} can be denoted as $W_l = \{W^1_l, \ldots, W^M_l\}$, where W^m_l is the filter parameter of the m^{th} layer. Through a series of nonlinear filtering operations in the hidden layers, the CDBN transforms the original input to a convolutional, high-level representation.

Note that for an input y_i, each of the feature extractors F_{W_l} generates the representation $\hat{y}^l_i = F_{W_l}(y_i)$, attuned to capturing salient information particular to the intermediate dataset D_l it is trained on. Considering the CDBNs corresponding to all points on the interpolating path, an input image is further represented by concatenating all of the outputs from the feature extractors together. Detailedly, the path feature $\hat{Y}_i$ for the input y_i is computed as:

$$\begin{aligned} \hat{Y}_i &= [F_{W_1}(y_i), F_{W_2}(y_i), \ldots, F_{W_L}(y_i)] \\ &= [\hat{y}^1_i, \hat{y}^2_i, \ldots, \hat{y}^L_i]. \end{aligned} \tag{5}$$

This new path representation incorporates the smooth distribution shift recovered in the interpolating CDBN path into the signal space. It brings the source and target data into a domain-robust feature space, where the sample differences caused by domain changes are reduced.

2.3 Discriminant Analysis on Path Feature

Since CDBN is trained in an unsupervised mode, the representation learned above is independent of tasks. In order to identify features with good discrimination and suitable for cross-view recognition, we develop a discriminant analysis framework on the path feature. Specially, the deep convolutional models project face images into a high-dimensional feature space, where samples are liable to be linearly separable [11]. Based on this, we form the discriminant training by learning linear transforms on the convolutional path feature, aiming at obtaining good discrimination in the mapping space. The learning objective is formulated by compelling faces from different domains towards that of the same identity, which effectively enhances the discrimination of the fine individual faces in the linear mapping space.

Assume $T = \{X_i \in X \bigcup Y_i \in Y\}$, $1 \le i \le C$ be the path feature set containing C identities. $X = \{X_1, X_2, \ldots, X_C\}$ is the source data, where $X_i = \{x_{i,k} \in \mathbb{R}^{d_S}\}^{N_{X_i}}_{k=1}$ denotes the feature of the i^{th} person, and N_{X_i} is the sample number. $Y = \{Y_1, Y_2, \ldots, Y_C\}$ holds the corresponding target features set $Y_i = \{y_{i,j} \in \mathbb{R}^{d_T}\}^{N_{Y_i}}_{j=1}$ for each person i in X. d_S and d_T are the feature dimensions. The linear transforms for the source and target views are denoted as $F_S \in \mathbb{R}^{d' \times d_S}$ and $F_T \in \mathbb{R}^{d' \times d_T}$, where d' is the mapping dimension. To enhance the discrimination, we use the intra-class compactness regularization as the objective function:

$$J(\theta_S, \theta_T) = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N_{X_i}} \sum_{k=1}^{N_{Y_i}} \|F_S x_{i,j} - F_T y_{i,k}\|^2. \tag{6}$$

where N is the number of sample pairs. We solve this problem with a simply matrix derivation. Let $X = [X_1, \ldots, X_C]$ collect the source data of all the person, where $X_i = [x_{i,1}, \ldots, x_{i,N_{X_i}}] \in \mathbb{R}^{d_S \times N_{X_i}}$ is the feature of person i. Similar denotations are used for the target feature matrix Y. In addition, we set:

$$\bar{X}_i = [\bar{x}_{i,1}, .., \bar{x}_{i,N_{X_i}}], \ \bar{x}_{i,j} = [x_{i,j}, .., x_{i,j}] \in \mathbb{R}^{d_S \times N_{Y_i}}, \tag{7}$$

$$\bar{Y}_i = [Y_i, .., Y_i] \in \mathbb{R}^{d_T \times (N_{Y_i} \times N_{X_i})}. \tag{8}$$

Then, we cast the function in Eq. (6) into a simplified form:

$$\min_{F_s,F_t} J = \frac{1}{N}\|F_S\bar{X} - F_T\bar{Y}\|_F^2, \tag{9}$$

where $\|.\|_F^2$ stands for the Frobenius norm. The gradient descend algorithm is used for optimization, where the gradients $\{\partial J/\partial F_S, \partial J/\partial F_T\}$ are easy to compute. Using the denotations above, the complexity of gradient computation costs $O(d'D^2N_{\bar{X}})$, where d' is the mapping dimension, D is the feature dimension and $N_{\bar{X}} = \sum_{i=1}^{C} N_{X_i}N_{Y_i}$. N_{X_i} and N_{Y_i} are the i^{th}-class sample numbers in the source set and target set respectively. C is the total class number of the training set. Besides, it just needs dozens of iterations before the updating converges experimentally. From this, the optimization of linear transforms in Eq. (9) is very computationally efficient on a limited training set.

The proposed CDBN path model appears to be similar with DLID [9], because they both learn a deep-model based interpolating path between cross-domain data. However, they have several significant differences in the aspects of model structure and training strategy. Specially, the CDBN path shows superiority over DLID as follows: 1) Holding simpler sub-model structure and pre-training framework. CRBM joints the filtering and pooling in a unified framework and is pre-trained by CD algorithm [13] to directly optimize the filter parameters, while the F-R-N-P model in DLID cascades four separate components together and performs pre-training in a sparse-coding framework which optimizes a series of parameters of the filters, coding dictionary and coefficients alternately [10]. 2) Learning high-level visual features. CRBM learns semantically meaningful features such as edges and face parts by taking advantage of the implicit structure of face images, while the F-R-N-P in DLID is trained on randomly selected image patches and generally learns edges and inapparent structures [14]. 3) Developing more simple and efficient discriminant training strategy. The CDBN path learns discriminant linear transforms for the convolutional path features to explicitly enhance the discrimination of samples in the mapping space, while DLID supervised fine-tunes all the nonlinear convolutional models on the interpolating path where the number of parameters to be adjusted is rather large relative to the training set and the resulting models are very likely to overfit on a limited training set.

3. Experiment

3.1 Dataset and Experiment Setting

CMU Multi-PIE [15] dataset contains 337 subjects, recorded under various poses, illumination and expressions. Following the setting in [1], the first 231 subjects are used for training, and the rest for testing. We select 6 images with varying illuminations of each subject under seven views ($-45°, -30°, -15°, 0°, 15°, 35°, 45°$) as the evaluation dataset. In our experiments, the front view ($0°$) is used as the source domain D_S, and the target domain D_T corresponds to the other six respectively. For each pair of cross-view datasets, we consider three datasets to form the interpolating path: the source-view dataset D_S, the intermediate dataset D_M, and the target-view dataset D_T. The only intermediate dataset D_M consists of half of D_S and half of D_T. Corresponding, three CDBNs are respectively trained on the three datasets. For the CDBN on each dataset, we train 24 first layer filters, each 10×10 pixels, and 40 second layer filters, each 12×12. The pooling ratio C is empirically set as 3 and 2 for the two layers respectively. For the path feature obtained, we first apply PCA to reduce the dimension as 1000, and then employ the discriminant learning.

Table 1 Recognition results of different models under varying target views on CMU Multi-PIE dataset (%).

Models	$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$	Avg
Linear	68.1	80.2	90.6	97.0	81.8	71.1	81.5
CDBN–L^1	75.6	90.2	94.7	98.7	91.0	83.5	89.0
CDBN–L^2	76.9	92.0	96.4	99.4	92.9	84.9	90.4
CDBN Path–L^1	76.3	91.2	95.8	99.5	92.0	84.3	89.8
CDBN Path–L^2	**80.0**	**95.3**	**99.5**	**99.8**	**95.9**	**88.4**	**93.2**
CDBN Path–L^3	78.9	94.3	98.7	99.5	95.4	87.6	92.4

3.2 Results and Analysis

To investigate the effectiveness of our method, we present the accuracy for individual models in Table 1. (1) Baseline "Linear" model, where linear transforms are learned for the source and target views. (2) CDBN, where a CDBN is trained with the data from both the source and target views, ignoring the notion of interpolating path. The one-layer and two-layer versions of models are denoted by L^1 and L^2. (3) The CDBN Path model. As shown, the deep models generally achieve better results than the linear one, suggesting that the deep learning architecture exhibits more powerful ability to capture the nonlinear variations of face images. The comparison of CDBN and CDBN Path demonstrates the benefit of intermediate representation learning along the interpolating path. It also indicates that modeling the correlation of cross-view data imposes much significance to prompt the performance of cross-view recognition task. Furthermore, the two-layer models (L^2) generally perform better than one-layer versions (L^1). This is in agreement with existing deep learning literature and provides justification for learning deep hierarchical representations. We also experiment by setting the layer number as 3 for the CDBN Path model, denoted as L^3 in Table 1. The average accuracy results in 92.4%, indicating that a larger layer number doesn't help to improve the performance of the proposed method. Generally, networks with deeper layers have more powerful modeling ability but also more parameters to be adjusted. For a limited dataset, too deep layers make the model much too complicated for the current problem, which further leads the model overfitting on the small training set while generalizing badly on the test set. To trade off on our dataset, we set the layer number as 2 for the proposed model according to the results in Table 1.

Table 2 reports the comparison with state-of-art meth-

Table 2 Performance comparison on CMU Multi-PIE dataset (%).

Methods	−45°	−30°	−15°	15°	30°	45°	Avg
PLS	81.0	86.2	93.0	94.5	82.5	79.0	86.0
CDEF	70.6	89.9	98.8	99.1	94.2	77.7	88.4
Dictionary Path	82.0	91.0	96.0	97.0	92.0	85.0	90.5
DLID	79.4	92.0	98.4	98.7	93.4	83.7	90.9
CDBN Path	80.0	95.3	99.5	99.8	95.9	88.4	93.2

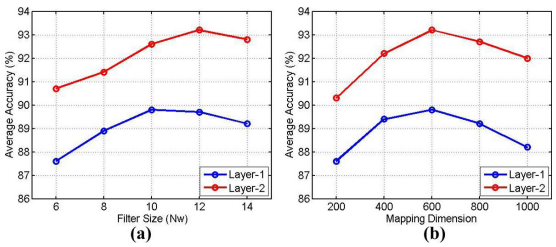

Fig. 4 (a) Performance with different filter sizes. (b) Performance with different mapping dimensions.

ods. We observe that classical linear models (CDEF [1], PLS [2]), which only consider single-view data for training, are heavily biased under cross views, and all the interpolating path methods improve upon them. Specially, the Dictionary Path method [16] learns representations by interpolating intermediate dictionaries between cross-view data. The comparison of Dictionary Path and CDBN Path demonstrates the advantage of deep nonlinear path model over the linear dictionary path model, suggesting the deep architecture models the source to target path information more effectively. Furthermore, we also verify the superior performance of CDBN Path over the similar DLID model [9] in Table 2. We experiment with the source code used in DLID, which was published by [14]. The CDBN Path performs better than DLID on face recognition for learning high-level structured features by exploiting the implicit structure of face images and potentially achieving better model discrimination with the well-designed linear discriminant learning stage. This result also provides further justification for the superiority of CDBN Path claimed in Sect. 2.

We also give a detailed analysis of the important parameters: N_W (the filter size of CRBM); the mapping dimension d'. First, we evaluate N_W of the first-layer CRBM by varying it from 6 to 14 in step of 2. The result is shown as the blue line (Layer-1) in Fig. 4 (a). As seen, just as N_W is around 10, the average accuracy reaches the peak. Fixing N_W as 10 in the first layer, we evaluate N_W for the second-layer CRBM in the same way, and show the result as the red line (Layer-2) in Fig. 4 (a). The filter size of 12 results in the highest accuracy. Actually, it is difficult to capture enough spatial information with small size filters, while overly large filter size result in severe over-smoothing of small details in the image. For the mapping dimension d', we evaluate by varying it from 200 to 1000 in step of 200. Figure 4 (b) shows the influence of d' to the one-layer (Layer-1) and two-layer (Layer-2) CDBN model. Similar variations occur for the two models. As seen, the accuracy benefits from increasing the mapping dimension. With too low dimension, performance drops for losing much discriminative information in the mapping operation. The best performance is obtained at $d' = 600$. A continued growth leads to a downward trend, because a high dimension leads the model overfitting on the small training set while generalizing badly on the test set.

4. Conclusion

We have developed a novel feature learning method to obtain high-level, domain-robust representations for cross-view face recognition. Convolutional Deep Belief Networks (CDBN) is introduced as the feature learning model. An CDBN based interpolating path is built to model the correlation of cross-view data. Moreover, the discriminant analysis is applied on the resulting features in order to gain more discriminative power. Comparative experiments demonstrate the proposal's high accuracy in cross-view recognition task.

References

[1] D. Lin and X. Tang, "Inter-modality face recognition," ECCV, pp.13–26, 2006.
[2] A. Sharma and D.W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," CVPR, pp.593–600, 2011.
[3] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," CVPR, pp.605–611, 2009.
[4] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," ECCV, pp.808–821, 2012.
[5] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," Conference on Data Mining and Data Warehouses (SiKDD 2010), pp.1–4, 2010.
[6] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol.313, no.5786, pp.504–507, 2006.
[7] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," ICML, pp.609–616, 2009.
[8] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," CVPR, pp.2855–2862, 2012.
[9] S. Chopra, S. Balakrishnan, and R. Gopalan, "Dlid: Deep learning for domain adaptation by interpolating between domains," ICML Workshop on Challenges in Representation Learning, 2013.
[10] K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," Tech. Rep. CBLL-TR-2008-12-01, 2008.
[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.8, pp.1798–1828, 2013.
[12] G.B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," CVPR, pp.2518–2525, 2012.
[13] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol.14, no.8, pp.1771–1800, 2002.
[14] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," ICCV, pp.2146–2153, 2009.
[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," Image and Vision Computing, vol.28, no.5, pp.807–813, 2010.
[16] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," CVPR, pp.692–699, 2013.