Conditional High-order Boltzmann Machine: A Supervised Learning Model for Relation Learning

Yan Huang[1] Wei Wang[1] Liang Wang[1,2]
[1]Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
[2]Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences
{yhuang, wangwei, wangliang}@nlpr.ia.ac.cn

Abstract

Relation learning is a fundamental operation in many computer vision tasks. Recently, high-order Boltzmann machine and its variants have exhibited the great power of modelling various data relation. However, most of them are unsupervised learning models which are not very discriminative and thus cannot server as a standalone solution to relation learning tasks. In this paper, we explore supervised learning algorithms and propose a new model named Conditional High-order Boltzmann Machine (CHBM), which can be directly used as a bilinear classifier to assign similarity scores for pairwise images. Then, to better deal with complex data relation, we propose a gated version of CHBM which untangles factors of variation by exploiting a set of latent variables to gate classification. We perform four-order tensor factorization for parameter reduction, and present two efficient supervised learning algorithms from the perspectives of being generative and discriminative, respectively. The experimental results of image transformation visualization, binary-way classification and face verification demonstrate that, by performing supervised learning, our models can greatly improve the performance.

1. Introduction

The goal of relation learning is to measure the similarity between samples, which is crucial to many retrieval, classification and verification tasks. To deal with that, in the past few years, researchers adapted the desired similarity measure to the form of metric and proposed various metric learning methods [9, 11, 42]. However, the metric assumption is insufficient to cover the diversity of data relation in the real world [5]. Recently, High-order Boltzmann Machine (HBM) [33] as a powerful relation learning model, has been applied to a range of tasks, e.g., analogy making [27], face verification [37], action recognition [39] and motion estimation [40].

The learning algorithms of HBM can be categorized into two main classes: conditional learning and joint learning. Given pairs of samples, the idea behind conditional learning is to use the latent variables to learn the conditional distribution of one sample given the other one [26, 27]. To overcome the difficulty that the conditional probability cannot be directly used to measure similarity in matching applications, joint learning alternatively learns the joint distribution over pairwise samples, where the joint probability can be used as a similarity score [37].

Both conditional learning and joint learning are performed in an unsupervised way, i.e., without using any relational labels, which is less discriminative for relation learning tasks. Taking face verification as an example, the goal is to assign a binary relational class (0 for "mismatched" and 1 for "matched") to a given pair of facial images. For this kind of binary classification problem, most HBM-based models just use the matched pairs of samples during training but ignore the mismatched ones [37, 19]. As a result, the learned models only impose constraints for the intra-label compactness but provide no guarantee for the inter-label separability, which is thus suboptimal for the discriminative tasks. In fact, the modelling of separability has been extensively studies by other similarity learning and metric learning methods such as [5, 9].

In this paper, to perform fully supervised learning and take the inter-label separability into consideration, we propose the Conditional High-order Boltzmann Machine (CHBM) which connects relational class labels to pairwise inputs with multiplicative interactions. The model can be regarded as a bilinear classifier for similarity, where the underlying assumption is that data relation can be linearly separated, and the probabilities of binary classes can be directly

inferred from the inputs. To better deal with very complex data relation, we propose an extended model called Gated CHBM, which makes no assumption about the data relation, but employs a set of latent variables to gate classification. The latent variables denote the untangled "environment" factor from the "class" factor, with the goal to explain the within-class variance.

Further to reduce cubicly many parameters produced by the multiplicative interactions, we propose four-order tensor factorization which approximates a four-order parameter tensor with four matrices. Then, we develop two supervised learning algorithms: 1) Generative learning optimizes the joint log-likelihood with stochastic gradient descent, where intractable gradients are efficiently approximated by a four-way version of Contrastive Divergence [14]. 2) Discriminative learning aims to optimize the conditional log-likelihood, where exact gradients can be directly computed. Afterwards, we demonstrate the effectiveness of our methods by applying them to the tasks of image transformation visualization, binary-way classification and face verification.

Our contributions can be summarized as follows. 1) We introduce supervised relational labels into conventional HBM with multiplicative interactions, and develop several effective supervised learning algorithms for relation learning. 2) To the best of our knowledge, we are the first to demonstrate the effectiveness of untangling factors of variation in the context of data relation. 3) Four-way Contrastive Divergence and four-order tensor factorization are explored for gradient approximation and parameter reduction, respectively.

2. Related Work

Our methods are closely related to the literature which uses "mapping units" [13] to learn data relation, especially the models based on High-order Boltzmann Machine (HBM) [33].

Gated Boltzmann Machine (GBM) [26, 27] is able to model image transformations by predicting one image conditioned on the other. But in such conditional learning, the conditional probability cannot be used to measure the similarity in matching tasks, because the probability is normalized with an unknown constant. To overcome this problem, MorphBM [37] learns the joint distribution over pairwise inputs, and directly uses the joint probability as a similarity score.

In contrast to the unsupervised conditional or joint learning, our models incorporate relational class labels to perform supervised learning. In particular, our CHBM replaces the latent variables of HBM with two "one-hot" encoded relational class variables. It should be noted that the latent variables of the Gated CHBM and those of HBM are fundamentally different, which denote an untangled "environment" factor and multiple tangled factors of variation, respectively. Our models also differ from the supervised learning model ClassRBM [21], which is mainly proposed for modelling data content but not data relation.

The proposed Gated CHBM is related to some RBM-based models which consider to untangle factors of variation. Gated softmax model [28] is a log-bilinear model, where the class probabilities are computed by multiplicatively integrating inputs with binary "style" features. Factored CRBM [40] employs a set of real-valued motion stylistic features to gate human motion analysis. With multi-way multiplicative interactions, disentangling RBM [32] untangles factors of variation from image content. Intrinsically different from these models above on data content, Gated CHBM untangles the factors of class and environment in the context of data relation.

3. Exploiting Relational Labels

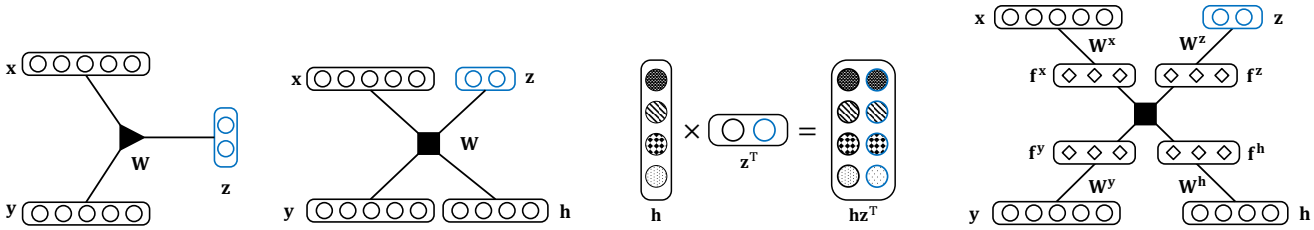
The task of relation learning can be formulated as follows: given a set of training data $\mathcal{D} = \{\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{z}^\alpha\}_{\alpha=1,\cdots,N}$, where α is the data index, $\mathbf{x}^\alpha$ and $\mathbf{y}^\alpha$ are a pair of input samples, and $\mathbf{z}^\alpha$ is the groundtruth relational class label, i.e., 0 for "mismatched" and 1 for "matched", the goal is to learn the projection from pairwise samples to relational classes.

To achieve this goal, we propose the Conditional High-order Boltzmann Machine (CHBM) as shown in Figure 1 (a). The model is an undirected graphical model which is composed of two sets of observed variables $\mathbf{x} = \{x_i\}_{i\in I}$ and $\mathbf{y} = \{y_j\}_{j\in J}$, and a set of class variables $\mathbf{z} = \{z_t\}_{t\in\{1,2\}}$. Here we assume $\mathbf{x}$ and $\mathbf{y}$ are binary-valued, i.e., $\mathbf{x} \in \{0,1\}^I$, $\mathbf{y} \in \{0,1\}^J$, the model can be easily generalized to handle real-valued inputs [43]. The two units z_1 and z_2 represent the probabilities of $\mathbf{x}$ and $\mathbf{y}$ are matched ($\mathbf{x} \sim \mathbf{y}$) or mismatched ($\mathbf{x} \nsim \mathbf{y}$), respectively. Since these two classes are mutually exclusive, the representation of them is "one-hot":

$$\begin{cases} \text{if } z_1 = 1,\ z_2 = 0, & \mathbf{x} \sim \mathbf{y} \\ \text{if } z_1 = 0,\ z_2 = 1, & \mathbf{x} \nsim \mathbf{y} \end{cases}$$

z_1 can be used as a real-value similarity measure in more general cases while z_2 can be used to measure dissimilarity. When more than two types of data relation is given, we can accordingly use more variables in the layer $\mathbf{z}$.

To perform content-independent similarity learning [25], the model uses two-way multiplicative interactions $\mathbf{x}\mathbf{y}^T$ between $\mathbf{x}$ and $\mathbf{y}$. Each element $x_i y_j$ can be regarded as an AND-gate which detects the correspondence between variables x_i and y_j. To directly model the projection from the detected correspondences to relational classes, the connections among $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are three-way multiplicative interactions, denoted by a three-order weight tensor $\mathbf{W} =$



(a) CHBM (b) Gated CHBM (c) multiplicative interactions (d) factored Gated CHBM

Figure 1. (a) and (b) are Conditional High-order Boltzmann Machine (CHBM) and Gated CHBM, respectively. (c) is the illustration of multiplicative interactions between class variables **z** and environment variables **h**. (d) is factored Gated CHBM. $\mathbf{f}^{\mathbf{x}}$, $\mathbf{f}^{\mathbf{y}}$, $\mathbf{f}^{\mathbf{h}}$ and $\mathbf{f}^{\mathbf{z}}$ are filter responses of **x**, **y**, **h** and **z**, respectively.

$\{W_{ijt}\}_{i\in I,j\in J,t\in\{1,2\}}$. Each weight W_{ijt} is associated with a triplet of variables $\{x_i, y_j, z_t\}$. Similar to Restricted Boltzmann Machines (RBM) [15], there is no internal connection among variables within each layer.

The energy function of the model is defined as follows:

$$E(\mathbf{x},\mathbf{y},\mathbf{z}) = -\sum_{ijt} W_{ijt}x_i y_j z_t - \mathbf{a}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} - \mathbf{d}^T\mathbf{z} \quad (1)$$

where **a**, **b** and **d** are biases of **x**, **y** and **z**, respectively. Based on the energy function, the joint distribution over all the variables is:

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{1}{Z}e^{-E(\mathbf{x},\mathbf{y},\mathbf{z})} \quad (2)$$

where $Z = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} e^{-E(\mathbf{x},\mathbf{y},\mathbf{z})}$ is a partition function for normalization.

During testing, we are particularly interested in predicting the relational class **z** given inputs **x** and **y**, where the classification decision is made by $\arg_t \max p(z_t|\mathbf{x},\mathbf{y})$:

$$p(z_t|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y},z_t)}{\sum_{t*} p(\mathbf{x},\mathbf{y},z_{t*})} = \frac{e^{\sum_{ij} W_{ijt}x_i y_j + d_t}}{\sum_{t*} e^{\sum_{ij} W_{ijt*}x_i y_j + d_{t*}}} \quad (3)$$

We can observe that, the model establishes a *log-bilinear* relation between pairwise inputs and relational classes. Specifically, the probability of each class is obtained by exponentiating and normalizing a class-specific bilinear score function $\sum_{ij} W_{ijt}x_i y_j + d_t$. Note that the score function is also a linear function of the detected correspondence $x_i y_j$, which potentially assumes that the data relation can be linearly separated by hyperplanes.

The model can be discriminatively learned by minimizing the conditional log-likelihood:

$$\mathcal{L} = -\sum_{\alpha} \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha,\mathbf{y}^\alpha) \quad (4)$$

over all the training data via stochastic gradient descent. The exact gradient of $\log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha,\mathbf{y}^\alpha)$ with respect to each model parameter $\theta \in \mathbf{W}$ is:

$$\frac{\partial \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha,\mathbf{y}^\alpha)}{\partial\theta} = \frac{\partial M_t^\alpha}{\partial\theta} - \sum_{t*} p(z_{t*}^\alpha|\mathbf{x}^\alpha,\mathbf{y}^\alpha)\frac{\partial M_{t*}^\alpha}{\partial\theta} \quad (5)$$

where $M_t^\alpha = \sum_{ij} W_{ijt}x_i^\alpha y_j^\alpha + d_t$.

4. Untangling Factors of Variation

As we know, data relation is composed of various factors of variation. For example, the relation of a pair of facial images depends on the factors of identity, expression and illumination. In previous work, the data relation is generally categorized into two classes in terms of matched or mismatched. In such way, the model only considers the class-related factor, but ignores other environmental ones. For example, in the tasks of face verification and face expression recognition, previous models only focus on modelling the factors of identity and expression, respectively, and ignore other environmental factors such as illumination and head pose. In the following, we propose Gated CHBM which aims to untangle factors of variation for data relation.

4.1. Model Description

The proposed Gated CHBM is illustrated in Figure 1 (b), which consists of two sets of observed variables **x** and **y**, a set of class variables **z**, and an additional set of latent variables $\mathbf{h} = \{h_k\}_{k\in K}$. Variables **z** and **h** are used to denote two factors of variation, namely "class" and "environment", respectively.

To untangle the two factors of class and environment, it is necessary to use multiplicative interactions between variables **z** and **h**. As shown in Figure 1 (c), the outer product $\mathbf{hz}^T$ produces eight environment-related subclasses, each of which is a free combination of the class and environment factors. When given an instantiation of the environment variables, the model actually performs an environment-free classification. The final classification decision can be obtained by marginalizing over the environment variables. As a result, the connections among **x**, **y**, **h** and **z** are four-way multiplicative interactions, denoted by a four-order weight tensor $\mathbf{W} = \{W_{ijkt}\}_{i\in I,j\in J,k\in K,t\in\{1,2\}}$.

The energy function of Gated CHBM is defined as:

$$E(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{z}) = -\sum_{ijkt} W_{ijkt}x_i y_j h_k z_t - \mathbf{a}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{t} \quad (6)$$

where **a**, **b**, **c** and **d** are biases of **x**, **y**, **h** and **z**, respectively. Then, we can obtain the joint distribution over all the variables $p(\mathbf{x}, \mathbf{y}, \mathbf{h}, \mathbf{z})$ similar to Equation 2.

Inheriting the conditionally independent property from RBM, we can perform tractable inferences as follows:

$$p(\mathbf{x}|\mathbf{y}, \mathbf{h}, \mathbf{z}) = \prod\nolimits_i \sigma(a_i + \sum\nolimits_{jkt} W_{ijkt} y_j h_k z_t) \quad (7)$$

$$p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \mathbf{z}) = \prod\nolimits_j \sigma(b_j + \sum\nolimits_{ikt} W_{ijkt} x_i h_k z_t) \quad (8)$$

$$p(\mathbf{h}|\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod\nolimits_k \sigma(c_k + \sum\nolimits_{ijt} W_{ijkt} x_i y_j z_t) \quad (9)$$

$$p(z_t|\mathbf{x}, \mathbf{y}, \mathbf{h}) = \frac{e^{d_t + \sum_{ijkt} W_{ijkt} x_i y_j h_k}}{\sum_{t*} e^{d_{t*} + \sum_{ijkt*} W_{ijkt*} x_i y_j h_k}} \quad (10)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.

During testing, before assigning the two general classes to pairwise inputs, the model first infers the probabilities of the environment-related subclasses (in Figure 1 (c)):

$$p(z_t, \mathbf{h}|\mathbf{x}, \mathbf{y}) = \frac{e^{\sum_{ijk} W_{ijkt} x_i y_j h_k + \sum_k c_k h_k + d_t z_t}}{\sum_{t*,\mathbf{h}} e^{\sum_{ijk} W_{ijkt*} x_i y_j h_k + \sum_k c_k h_k + d_{t*} z_{t*}}} \quad (11)$$

Note that the model is actually a *log-trilinear* model, since the probability of each subclass is computed by exponentiating and normalizing the trilinear score function $\sum_{ijk} W_{ijkt} x_i y_j h_k$. Then, we can compute $p(z_t|\mathbf{x}, \mathbf{y})$ by marginalizing over the latent variables **h**:

$$p(z_t|\mathbf{x}, \mathbf{y}) = \frac{e^{d_t + \sum_k \log(1 + e^{c_k + \sum_{ij} W_{ijkt} x_i y_j})}}{\sum_{t* \in \{1,2\}} e^{d_{t*} + \sum_k \log(1 + e^{c_k + \sum_{ij} W_{ijkt*} x_i y_j})}} \quad (12)$$

The model can also be interpreted as a mixture model. Each environment variable h_k blends in a three-dimensional slice $\mathbf{W}_{..k.}$, corresponding to an environment-specific CHBM. Since the model integrates out totally 2^K possible combinations of the K environment variables, it is exactly the same as a mixture of 2^K CHBMs. It should be noted that, in contrast to CHBM, Gated CHBM makes no assumption about the specific form of the separation boundary, but just uses a set of latent variables to multiplicatively gate classification.

4.2. Four-order Tensor Factorization

To reduce the large number of parameters in the four-order weight tensor **W**, we perform a *four-order tensor factorization* which factors the tensor into four weight matrices $\mathbf{W^x} = \{W^{\mathbf{x}}_{if}\}_{i \in I, f \in F}$, $\mathbf{W^y} = \{W^{\mathbf{y}}_{jf}\}_{j \in J, f \in F}$, $\mathbf{W^h} = \{W^{\mathbf{h}}_{kf}\}_{k \in K, f \in F}$ and $\mathbf{W^z} = \{W^{\mathbf{Z}}_{tf}\}_{t \in \{1,2\}, f \in F}$, where F is the number of hidden states. In detail, each element W_{ijkt} is approximated using a four-way inner product:

$$W_{ijkt} = \sum\nolimits_{f=1}^{F} W^{\mathbf{x}}_{if} W^{\mathbf{y}}_{jf} W^{\mathbf{h}}_{kf} W^{\mathbf{t}}_{tf} \quad (13)$$

Algorithm 1 The generative learning of Gated CHBM.

Input: training data $\{\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{z}^\alpha\}$, learning rate λ
Notation: $a \leftarrow b$: setting a as value b
$\widehat{a} \sim a$: sampling $\widehat{a}$ from a
// M update iterations
for $m = 1$ to M **do**
  // Positive phase
  $\mathbf{x}^{(0)} \leftarrow \mathbf{x}^\alpha$, $\mathbf{y}^{(0)} \leftarrow \mathbf{y}^\alpha$, $\mathbf{z}^{(0)} \leftarrow \mathbf{z}^\alpha$,
  $\mathbf{h}^{(0)} \leftarrow p(\mathbf{h}|\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \mathbf{z}^{(0)})$
  // Negative phase
  $\widehat{\mathbf{h}^{(0)}} \sim p(\mathbf{h}|\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \mathbf{z}^{(0)})$
  $s \sim$ Uniform(0, 0.6) // Six sampling cases
  if $0 \leq s < 0.1$ **do**
    $\mathbf{x}^{(1)} \sim p(\mathbf{x}|\mathbf{y}^{(0)}, \widehat{\mathbf{h}^{(0)}}, \mathbf{z}^{(0)})$,
    $\mathbf{y}^{(1)} \sim p(\mathbf{y}|\mathbf{x}^{(1)}, \widehat{\mathbf{h}^{(0)}}, \mathbf{z}^{(0)})$,
    $\mathbf{z}^{(1)} \sim p(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \widehat{\mathbf{h}^{(0)}})$
  else if $0.1 \leq s < 0.2$ **do**
    . . . . . .
  end if
  $\mathbf{h}^{(1)} \leftarrow p(\mathbf{h}|\mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \mathbf{z}^{(1)})$
  // Update parameters
  for $\theta \in \Theta$ **do**
    $\Delta\theta \leftarrow \frac{\partial}{\partial\theta} E(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \mathbf{h}^{(0)}, \mathbf{z}^{(0)})$
      $- \frac{\partial}{\partial\theta} E(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \mathbf{h}^{(1)}, \mathbf{z}^{(1)})$
    $\theta \leftarrow \theta - \lambda\Delta\theta$
  end for
end for

The factored Gated CHBM is illustrated in Figure 1 (d), whose energy function can be obtained by plugging Equation 13 in Equation 6:

$$E_f(\mathbf{x}, \mathbf{y}, \mathbf{h}, \mathbf{z}) = -\sum\nolimits_{fijkt} (W^{\mathbf{x}}_{if} x_i)(W^{\mathbf{y}}_{jf} y_j)(W^{\mathbf{h}}_{kf} h_k)(W^{\mathbf{z}}_{tf} z_t) - \mathbf{a}^T\mathbf{x} - \mathbf{b}^T\mathbf{y} - \mathbf{c}^T\mathbf{h} - \mathbf{d}^T\mathbf{t} \quad (14)$$

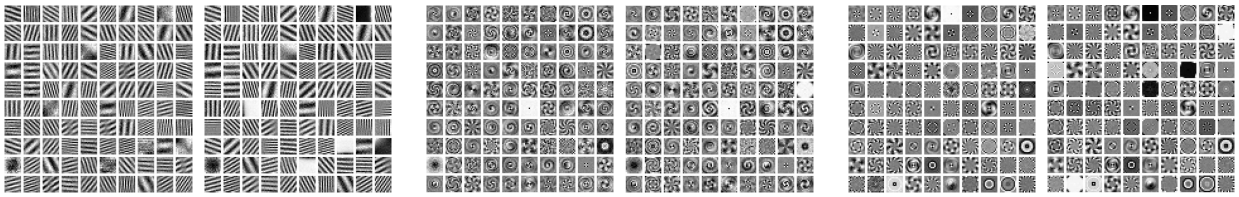
where the energy first fits **x**, **y**, **h** and **z** to F filters $\mathbf{W^x}$, $\mathbf{W^y}$, $\mathbf{W^h}$ and $\mathbf{W^z}$, respectively, and then sums over products of corresponding filter responses. The energy will assign small values when the filter responses tend to match well. Such filter matching amounts to finding suitable filters that can well explain the data relation.

4.3. Learning

Generative Learning: The learning procedure aims to minimize the negative joint log-likelihood:

$$\mathcal{L}_{gen} = -\sum\nolimits_\alpha \log p(\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{z}^\alpha) \quad (15)$$

with stochastic gradient descent. The exact gradient with



(a) translation-randomness (b) rotation-randomness (c) translation-rotation-randomness

Figure 2. Visualization of learned filters by Gated CHBM on the synthetic dataset.

respect to a model parameter $\theta \in \mathbf{W}$ is:

$$\frac{\partial \log p(\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{z}^\alpha)}{\partial \theta} = - E_{\mathbf{h}|\mathbf{x}^\alpha,\mathbf{y}^\alpha,\mathbf{z}^\alpha}[\frac{\partial}{\partial \theta} E(\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{h}, \mathbf{z}^\alpha)] + E_{\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{z}}[\frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{y}, \mathbf{h}, \mathbf{z})] \tag{16}$$

where the model expectation (the second term on the right side) is intractable. So we develop a *four-way Contrastive Divergence* to approximate it. In particular, we generate new samples by performing Gibbs sampling alternatively from one of the four distributions $p(\mathbf{x}|\mathbf{y}, \mathbf{h}, \mathbf{z})$, $p(\mathbf{y}|\mathbf{x}, \mathbf{h}, \mathbf{z})$, $p(\mathbf{h}|\mathbf{x}, \mathbf{y}, \mathbf{z})$, $p(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathbf{h})$. Different from the bi-partite RBM, Gated CHBM is a quad-partite model which has to visit four sets of variables during one-step sampling, where we have to decide which set to sample first. To reduce the bias caused by the order, the sampling is performed in a random order. The detailed learning procedure with one-step Gibbs sampling is shown in Algorithm 1.

Discriminative Learning: We utilize the a more discriminative objective [21] based on the conditional log-likelihood:

$$\mathcal{L}_{dis} = -\sum\nolimits_\alpha \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha, \mathbf{y}^\alpha) \tag{17}$$

For the unfactored Gated CHBM, the gradient of $\log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha, \mathbf{y}^\alpha)$ with respect to the model parameter $\theta \in \mathbf{W}$ can be computed exactly:

$$\frac{\partial \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha, \mathbf{y}^\alpha)}{\partial \theta} = \sum\nolimits_k \sigma(M^\alpha_{kt}) \frac{\partial M^\alpha_{kt}}{\partial \theta} - \sum\nolimits_{kt*} \sigma(M^\alpha_{kt*}) p(z^\alpha_{t*}|\mathbf{x}^\alpha, \mathbf{y}^\alpha) \frac{\partial M^\alpha_{kt*}}{\partial \theta} \tag{18}$$

where $M^\alpha_{kt} = c_k + \sum_{ij} W_{ijkt} x^\alpha_i y^\alpha_j z^\alpha_t$. Note that gradients with respect to biases $\mathbf{a}$ and $\mathbf{b}$ are 0 since they are eliminated in $p(\mathbf{y}|\mathbf{x})$. Then, for the factored model, we can compute the gradient with respect to $W^{\mathbf{x}}_{if}$ using the chain rule:

$$\frac{\partial \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha, \mathbf{y}^\alpha)}{\partial W^{\mathbf{x}}_{if}} = \sum\nolimits_{jkt} \frac{\partial \log p(\mathbf{z}^\alpha|\mathbf{x}^\alpha, \mathbf{y}^\alpha)}{\partial W_{ijkt}} \frac{\partial W_{ijkt}}{\partial W^{\mathbf{x}}_{if}} \tag{19}$$

where we can use Equations 18 and 13 to compute the two terms on the right side.

We experimentally find that, when performing discriminative learning with random parameter initialization, the model tends to be stuck in some local optima. To overcome this issue, we use a two-phrase learning algorithm: 1) pre-training the model with generative learning for a few iterations[1] to obtain better initializations [16], and 2) fine-tuning the parameters with discriminative learning.

5. Experiments

To verify the effectiveness of the proposed models, we perform two experiments including image transformation visualization, binary-way classification and face verification.

5.1. Image Transformation Visualization

Since the Gated CHBM are explained as filter-matching, we want to test whether the model can indeed learn some meaningful filters. The experimental dataset contains synthetic random dot images, each of which has a size of 13×13[2]. Each pixel is selected to be white with the probability of 0.1. Note that each image itself has no content, but pairwise images can belong to one of three transformations including translation, rotation and randomness. We generate 10,000 pairs of images for each transformation. When generating pairwise translated images, the translated steps are randomly sampled from the interval $[-3, 3]$ in both vertical and horizontal directions. For rotation, the rotated angles are randomly sampled from $[0°, 359°]$. For randomness, the pairwise images contain no specific relation. Since Gated CHBM is a supervised learning model, we take the translated (rotated) and random pairs of images as samples of two relational classes. We use 20 pairs as a minibatch during each iteration, and set the numbers of hidden units and factors as 200 and 100, respectively[3].

The learned pairwise filters ($\mathbf{W}^{\mathbf{x}}$ and $\mathbf{W}^{\mathbf{y}}$) on translated and random pairwise images are shown in Figure 2 (a), where the filters are similar to Fourier basis, representing

[1] In our experiments, we observe that 30 iterations are generally sufficient.

[2] Our model can be scaled to larger images with a size of 32×32, their visualization results are similar.

[3] In fact, varying these hyperparameters does not have significant impact on visualization.

Table 1. Accuracies of binary-way classification by all the compared methods on the MNIST-variant datasets.

Method	$basic$	rot
Cosine	69.69 ± 0.43	56.22 ± 1.61
ITML [9]	80.44 ± 0.09	60.61 ± 1.07
Gated RBM [27]	73.63 ± 0.34	67.06 ± 0.81
MorphBM [37]	91.48 ± 0.04	79.46 ± 0.11
CHBM	93.90 ± 0.09	81.12 ± 0.05
Gated CHBM-gen	93.13 ± 0.09	80.31 ± 0.07
Gated CHBM-dis	95.01 ± 0.05	83.74 ± 0.04

translation with different directions and steps. Figure 2 (b) shows pairwise filters learned on rotated and random pairs which are similar to a log-polar version of Fourier basis, containing circular and spiral patterns. There are also some random filters in the two figures, which account for the random transformation. Moreover, there exists the quadrature phase difference between pairwise filters, i.e., the phrase difference is about $90°$. In contrast to unsupervised HBM, the filter matching of our model is under the supervision of class labels. By assigning small energy to well-matched filter responses (in Equation 6), the model can explicitly establish the dependency relation between the learned filters and class labels.

We also take the three transformations as three classes and re-train the model. The learned filters are shown in Figure 2 (c). Compared with Figures 2 (a) and (b), there are more class-shared filters which exhibit fine black and white granular and center symmetric patterns. These filters can alternatively account for each of the three transformations.

5.2. Binary-way Classification

To study the capacity of handling various factors of variation, we apply our models to the task of binary-way classification, whose goal is to measure the similarities between pairwise samples, and assign binary relational classes ("matched" or "mismatched") to them. Binary-way classification can be regarded as a preliminary procedure for the task of invariant recognition [37, 5]. The experimental datasets are two variants of MNIST [22], including $basic$ and rot, where the images contain different factors of variation such as hand writing style and rotation angle. In each of the two datasets, we randomly generate 20,000 (20,000), 2,000 (2,000), 10,000 (10,000) matched (mismatched) pairs of images for training, validation and testing from the corresponding sets, respectively, and repeat for five times. Note that two images are treated as matched as long as they belongs to the same digital class.

We compare our models with four distance metric learning or similarity learning methods, including Cosine similarity, ITML [9], Gated RBM [27] and MorphBM [37]. Some hyperparameters such as the number of latent variables and learning rate are all selected based on the binary-way classification accuracy on the validation set. For the Gated CHBM, we study its performance under two settings in terms of generative and discriminative learning, denoted by suffixes "-gen" and "-dis", respectively.

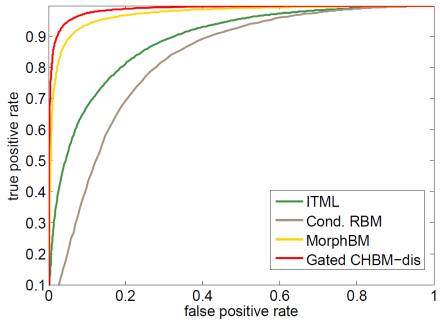


Figure 3. ROC curves of binary-way classification on the $basic$ dataset.

The accuracies of all the compared methods are shown in Table 1, from which we can see that our models consistently outperform all the compared methods on the two datasets. In particular, compared with the unsupervised learning methods Gated RBM and MorphBM, CHBM and Gated CHBM greatly improve the classification performance, which demonstrates the effectiveness of exploiting relational class labels for supervised learning. For Gated CHBM, due to the usage of a discriminative objective, discriminative learning can always yield higher accuracies than generative learning. With either generative or discriminative learning, Gated CHBM performs better than CHBM, which results from the fact that Gated CHBM makes the less assumption about the decision boundary of classification, and is able to leverage potential resources to promote the accuracies.

We take the output probabilities of the matched class as similarity scores, and draw the Receiver Operating Characteristic (ROC) curves in Figure 3. We can observe that, Gated CHBM-dis presents the best visualization among all the methods, which is in consistent with the classification accuracies in Table 1.

5.3. Face Verification

In this section, we will apply our models to a more challenging task, namely face verification. The goal of face verification is to decide whether a given pair of facial images are matched or not.

In our experiment, we will use two facial datasets: 1) LFW [20]. The LFW dataset consists of totally 13,233 facial images from 5,749 different individuals. Among all of them, 1,680 individuals have at least 2 images while the rest have only a single image. Since all the images are collected from the Internet, there exists very large intra-person variation. 2) Multi-PIE [10]. The images of this dataset come

from 337 different individuals, which are captured under various view points, illumination conditions and facial expressions.

Due to the large intra-person variation, directly using relation learning models to handle raw facial images is very hard as discussed in [19]. So we first exploit some powerful hand-crafted descriptors such as LBP [1] to extract robust facial features, and then perform relation learning based on the obtained representations. The procedure of feature extraction includes localizing dense facial landmarks [3], extracting multi-scale[4] features [7] around each landmark, utilizing PCA for dimensionality reduction[5] on the concatenated high-dimensional features, and performing intra-PCA [4, 41] for intra-personal variation reduction.

Recently, the highest accuracy on the LFW dataset has reached over 99% by [36], in which the usage of large-scale labeled training data outside of LFW plays a significant role. However, it should be noted that, here we only take face verification as a case study to validate the effectiveness of our methods for relation learning, rather than vastly boost the performance. Therefore, in this experiment, we do not use any labeled outside data during training, and only focus on the dataset itself under two commonly used protocols [20]: 1) restricted protocol, label-free outside data and 2) unrestricted protocol, label-free outside data.

5.3.1 Restricted Protocol

Here we closely follow the public restricted protocol on the LFW dataset, which splits all the data into ten folds and performs ten-fold cross validation. Note that since the individual name of each facial image is unknown, we can only use the restricted number of image pairs for training. Similarly, on the Multi-PIE dataset, for each of ten times cross validation, we randomly select 49 identities for testing and the rest for training, and generate 5,400 and 600 pairs for training and testing, respectively.

In addition to ITML [9], Gated RBM [27] and MorphBM [37], we also compare our models with Sub-SML [4] which is a state-of-the-art method for face verification. We use the same facial representations for all the methods and the corresponding accuracies on the two datasets are illustrated in Table 2. As we can see, all our CHBMs can achieve better performance than Gated RBM and MorphBM, which demonstrates their discrimination of exploiting relational labels. Gated CHBM consistently outperforms CHBM[6] on the two datasets, which indicates that both learning relational features and untangling factors of variation are useful for modelling data relation.

Table 2. Accuracies of face verification by all the compared methods on the LFW and Multi-PIE datasets, under the restricted protocol (all the methods use the same facial representations as inputs).

Method	LFW	Multi-PIE
ITML [9]	77.90 ± 3.55	88.46 ± 4.15
Sub-SML [4]	86.93 ± 4.90	91.20 ± 3.66
Gated RBM [27]	82.45 ± 2.85	92.66 ± 1.28
MorphBM [37]	85.20 ± 1.51	93.58 ± 0.44
CHBM	88.90 ± 0.91	94.55 ± 0.95
Gated CHBM-gen	90.21 ± 1.25	94.75 ± 1.93
Gated CHBM-dis	89.55 ± 0.96	96.10 ± 0.53

Table 3. Accuracies of face verification by state-of-the-art methods on the LFW dataset, under the restricted protocol (the compared results are directly cited from already published papers.). Methods marked with $*$ are published after the submission of this paper.

Method	Accuracy
PAF [44]	87.77 ± 0.51
Convolutional DBN [23]	87.77 ± 0.62
CSML [29]	88.00 ± 0.37
HTBIF [31]	88.13 ± 0.58
SFRD+PMML [8]	89.35 ± 0.50
LM3L [18]	89.57 ± 1.53
Sub-SML [4]	89.73 ± 0.38
DDML [17]	90.68 ± 1.41
VMRS [2]	91.10 ± 0.59
Sub-SML + Hybrid on LFW3D [12]*	91.65 ± 1.04
HPEN + HD-Gabor + DDML [45]*	92.80 ± 0.47
Ours (Gated CHBM-gen)	91.70 ± 0.98

On the LFW dataset, we follow the score combination strategy in [38, 6] to further improve the face verification accuracy. We first obtain two similarity scores of Gated CHBM-gen on LBP and SIFT descriptors, respectively, and then classify the concatenated similarity scores with a linear SVM. We compare the improved result with the state-of-the-art methods[7] on the LFW dataset in Table 3. Note that the results of the compared methods are from the LFW websites[8]. From the table, we can see that Gated CHBM-gen achieves 91.70% accuracy. We also present ROC curves in Figure 4 (a), which shows that our model can obtain higher true positive rates when false positive rates are low.

5.3.2 Unrestricted Protocol

In this protocol, the individual name of each image is available, so we can generate as many matched and mismatched pairs as desired. In our experiment, on the LFW dataset, we generate 15,000 matched and 15,000 mismatched pairs

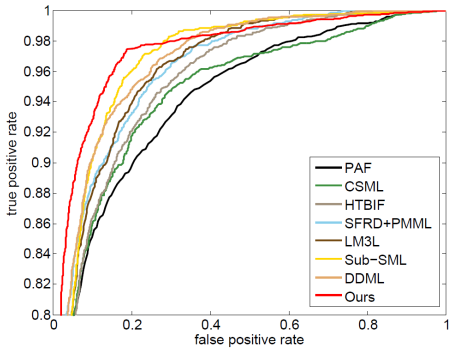
[4]The sizes of the image in each scale are [300,300], [212,212], [150,150], [106,106], [75,75].

[5]We vary the PCA dimensions from 100 to 2,000, but find it does not change the order of performance. In addition, most methods can achieve their best performance when the dimension is 400.

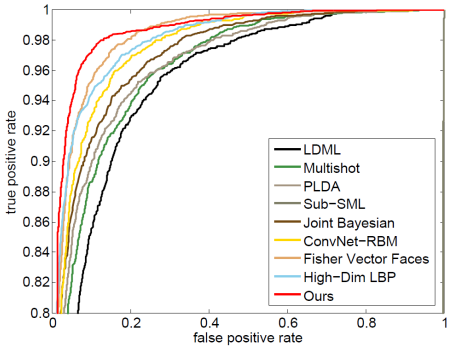
[6]To make a fair comparison with Gated CHBM, we perform a three-way tensor factorization for CHBM (similar to Equation 13).

[7]Without using large-scale labeled training data outside of LFW.

[8]*http://vis-www.cs.umass.edu/lfw/results.html*.



(a) Restricted protocol



(b) Unrestricted protocol

Figure 4. ROC curves by state-of-the-art methods on the LFW dataset, under restricted and unrestricted protocols.

Table 4. Accuracies of face verification by all the compared methods on the LFW and Multi-PIE datasets, under the unrestricted protocol (all the methods use the same facial representations as inputs).

Method	LFW	Multi-PIE
ITML [9]	87.73 ± 3.96	94.21 ± 1.42
LDML [11]	88.13 ± 2.88	95.38 ± 0.92
Sub-SML [4]	87.58 ± 4.64	93.63 ± 1.78
Gated RBM [27]	86.30 ± 2.59	94.75 ± 1.49
MorphBM [37]	89.95 ± 1.23	96.70 ± 0.54
CHBM	90.18 ± 1.15	96.01 ± 0.85
Gated CHBM-gen	91.06 ± 0.98	96.97 ± 0.71
Gated CHBM-dis	90.36 ± 1.41	97.95 ± 0.42

Table 5. Accuracies of face verification by state-of-the-art methods on the LFW dataset, under the unrestricted protocol (the compared results are directly cited from already published papers.). Methods marked with $*$ are published after the submission of this paper.

Method	Accuracy
LDML [11]	87.50 ± 0.40
Multishot [38]	89.50 ± 0.51
PLDA [24]	90.07 ± 0.51
Sub-SML [4]	90.75 ± 0.64
Joint Bayesian [6]	90.90 ± 1.48
ConvNet-RBM [35]	91.75 ± 0.48
VMRS [2]	92.05 ± 0.45
Fisher Vector Faces [34]	93.03 ± 1.05
MLBPH+MLPQH+MBSIFH [30]	93.03 ± 0.82
High-Dim LBP [7]	93.18 ± 1.07
HPEN + HD-Gabor + DDML [45]*	95.25 ± 0.36
Ours (Gated CHBM-gen)	93.73 ± 0.85

of images for each time training[9]. While on the Multi-PIE dataset, we generate 10,000 matched and 10,000 mismatched pairs of images for training, repeated for ten times.

Under this protocol, we additionally compare with another baseline LDML [11]. The recognition accuracies of all the compared methods under this protocol are shown in Table 4. We can find the overall gains in performance for all the methods when compared with Table 2, which results from the usage of more training pairs. In addition, our methods still surpass MorphBM by 1.11 % and 1.25 % on the two datasets, respectively. Note that the promotions are not so significant as those in Table 2, which indicates that our methods can make better use of limited training data to achieve more discriminate results.

On the LFW dataset, we further compare our best Gated CHBM-gen with the state-of-the-art methods (under the unrestricted protocol) in Table 5, and draw their ROC curves in Figure 4 (b). Similar to the restricted protocol, we use score combination to further improve the accuracy of Gated CHBM-gen to 93.73 %. From both table and figure, we can find that our method performs better than most state-of-the-art methods.

6. Conclusion

In this paper, to utilize relational labels for supervised relation learning, we have proposed a Conditional High-order Boltzmann Machine (CHBM), which is a log-bilinear classifier for data relation. We also have proposed an improved model as Gated CHBM which untangles factors of variation in the context of data relation. We have demonstrated the effectiveness of our methods by performing experiments of image transformation visualization, binary-way classification and face verification. In the future, we will apply our models to more relation learning tasks.

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China (61420106015, 61175003, 61202328, 61572504) and National Basic Research Program of China (2012CB316300).

[9] In fact, when the number of training pairs becomes larger than 30,000, the performance remains unchanged.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patternspatterns: Application to face recognition. *IEEE TPAMI*, 2006.

[2] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication face recognition. *ICCV*, 2013.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. *ICCV*, 2013.

[4] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. *ICCV*, 2013.

[5] S. Changpinyo, K. Liu, and F. Sha. Similarity component analysis. *NIPS*, 2013.

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. *ECCV*, 2012.

[7] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. *CVPR*, 2013.

[8] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. *CVPR*, 2013.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information theoretic metric learning. *ICML*, 2007.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 2010.

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. *ICCV*, 2009.

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. *CVPR*, 2015.

[13] G. E. Hinton. A parallel computation that assigns canonical object-based frames of reference. *IJCAI*, 1981.

[14] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[16] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[17] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. *CVPR*, 2014.

[18] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. *ACCV*, 2014.

[19] G. B. Huang and E. Learned-Miller. Learning class-specific image transformations with higher-order boltzmann machines. *CVPRW*, 2010.

[20] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *UMass Amherst Technical Report*, 2014.

[21] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. *ICML*, 2008.

[22] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. *ICML*, 2007.

[23] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML*, 2009.

[24] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince. Probabilistic models for inference about identity. *IEEE TPAMI*, 2012.

[25] R. Memisevic. Learning to relate images. *IEEE TPAMI*, 2013.

[26] R. Memisevic and G. Hinton. Unsupervised learning of image transformations. *CVPR*, 2007.

[27] R. Memisevic and G. Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 2010.

[28] R. Memisevic, C. Zach, G. Hinton, and M. Pollefeys. Gated softmax classification. *NIPS*, 2010.

[29] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. *ACCV*, 2010.

[30] A. Ouamane, B. Messaoud, A. Guessoum, A. Hadid, and M. Cheriet. Multi-scale multi-descriptor local binary features and exponential discriminant analysis for robust face authentication. *ICIP*, 2014.

[31] N. Pinto and D. Cox. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. *FG*, 2011.

[32] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. *ICML*, 2014.

[33] T. J. Sejnowski. Higher-order boltzmann machines. *Neural Networks for Computing*, 1986.

[34] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. *BMVC*, 2013.

[35] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. *ICCV*, 2013.

[36] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *NIPS*, 2014.

[37] J. Susskind, R. Memisevic, G. Hinton, and M. Pollefeys. Modeling the joint density of two images under a variety of transformations. *CVPR*, 2011.

[38] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. *BMVC*, 2009.

[39] G. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. *ECCV*, 2010.

[40] G. Taylor and G. Hinton. Factored conditional restricted boltzmann machines for modeling motion style. *ICML*, 2009.

[41] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE TPAMI*, 2004.

[42] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[43] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. *NIPS*, 2004.

[44] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. *CVPR*, 2013.

[45] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. *CVPR*, 2015.