Capturing Global Semantic Relationships for Facial Action Unit Recognition

Ziheng Wang[1] Yongqiang Li[2] Shangfei Wang[3] Qiang Ji[1]
[1]ECSE Department, Rensselaer Polytechnic Institute
[2]School of Electrical Engineering and Automation, Harbin Institute of Technology
[3]School of Computer Science and Technology, University of Science and Technology of China
{wangz10,liy23,jiq}@rpi.edu sfwang@ustc.edu.cn

Abstract

In this paper we tackle the problem of facial action unit (AU) recognition by exploiting the complex semantic relationships among AUs, which carry crucial top-down information yet have not been thoroughly exploited. Towards this goal, we build a hierarchical model that combines the bottom-level image features and the top-level AU relationships to jointly recognize AUs in a principled manner. The proposed model has two major advantages over existing methods. 1) Unlike methods that can only capture local pair-wise AU dependencies, our model is developed upon the restricted Boltzmann machine and therefore can exploit the global relationships among AUs. 2) Although AU relationships are influenced by many related factors such as facial expressions, these factors are generally ignored by the current methods. Our model, however, can successfully capture them to more accurately characterize the AU relationships. Efficient learning and inference algorithms of the proposed model are also developed. Experimental results on benchmark databases demonstrate the effectiveness of the proposed approach in modelling complex AU relationships as well as its superior AU recognition performance over existing approaches.

1. Introduction

The facial action units (AUs), as defined in the Facial Action Coding System (FACS) [4], refer to the local facial muscle actions. The AUs occur in different combinations and can lead to a huge variety of complex facial behaviors. Automatic analysis of these facial action units is of great importance in a wide range of fields such as behavioral understanding, video conference, affective computing, human-machine interface and so on.

Facial action units have been mainly treated as unrelated entities and recognized individually, based on either shape or appearance features. However, AUs are NOT independent from one another. On the one hand, facial action units generally do not occur alone and some combinations of action units are frequently observed. As illustrated in Figure 1, when you smile, you "raise cheek" and "pull lip corners". When you are surprised, you "stretch mouth", "raise brow" and "raise eye lid". These AUs tend to occur simultaneously to express different human emotions. On the other hand, some AUs must or must not be present at the same time due to the limitations of facial anatomy. For instance, "stretch mouth" prevents "raise cheek". Meanwhile, it is difficult to "wrinkle nose" without "raise cheek". While the presence or absence of each action unit can be inferred from the low-level facial shape and appearance changes, it is also significantly influenced by the states of the related action units. Therefore, an automatic facial action unit recognition system should not only rely on the low-level features, but also take advantage of the high-level semantic relationships among all the action units.

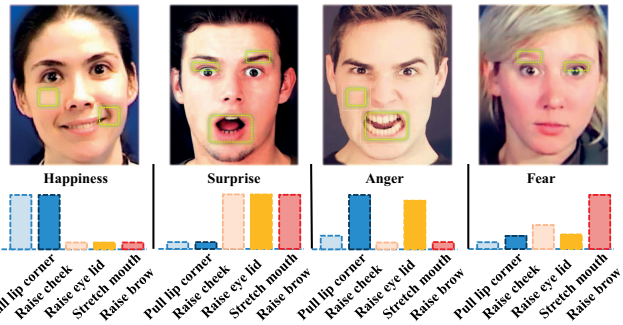

Figure 1: Examples showing: 1) some AU combination patterns are frequently observed. Each bar shows the intensity of presence for the corresponding action unit. Five AUs are illustrated in the images. 2) AU relationships depend on facial expressions. Consider "raise brow" and "stretch mouth". They tend to be both absent during happiness, both present during surprise, and either present during anger or fear.

Despite the importance, AU relationships are difficult to

capture and have not been fully exploited by the existing works. First, current models such as the Bayesian networks (BN) [20] are based on the first-order Markov assumption and therefore can only capture local, i.e. pairwise relationships between action units. Moreover, finding the optimal structure of a large AU network is difficult. An advanced AU recognition algorithm should efficiently and effectively capture not only local but also global dependencies among AUs. Second, the relationships among action units are influenced by many factors, including the expression, identity, age and gender of the subject. For instance the dependence between "stretch mouth" and "raise brow" is significantly affected by the human expression. As shown in Figure 1, these two action units are likely to be both absent during happiness, both present during surprise, and mutually exclusive during anger or fear. Current works ignore these factors, which could lead to incorrect estimation of the AU relationships.

Unlike a regular BN, the restricted Boltzmann machine (RBM) and its variants can model higher-order dependencies among random variables by introducing a layer of latent units. RBMs have been widely used for modeling complex joint distributions over structured variables such as image pixels [6]. This motivates us to propose a hierarchical model to simultaneously address all the above issues. Unlike [20], we introduce RBM to model the action units, thereby capturing not only local but also global AU dependencies. To the best of our knowledge, this is the first time RBM is used to model the action units. Furthermore, a 3-way RBM [14] is applied to incorporate the related factors that can affect AU relationships. Finally, the proposed model combines bottom-level image measurements with the top-level prior AU semantics in a principled manner, and we propose efficient algorithms for its learning and inference.

The remainder of this paper is organized as follows. Section 2 presents an overview of the related works. We briefly review the restricted Boltzmann machine in Section 3 and then introduce our algorithm in detail in Section 4. Experiments and discussions will be illustrated in Section 6. The paper is concluded in Section 7.

2. Related Works

A number of approaches have been proposed for facial action unit recognition and they generally involve two steps: bottom-level facial feature extraction and top-level AU classifier design.

Facial features: The facial features for AU recognition can be grouped into appearance features and geometric features. The appearance features capture the local or global appearance changes of the facial components. Widely used features in this category include the Haar feature [25], local binary patterns (LBP) [7], Gabor wavelets [2], the canonical appearance feature [11] and others. Geometric features represent the changes in the direction or magnitude of the skin surface and salient facial points caused by facial muscular activities. These geometric changes can be measured via dense optical flows [10] or the displacement of facial feature points [23, 11], etc.

AU classifiers. With the features, action units can be classified using two types of approaches. The first type of approaches treats action units as unrelated entities. The action units are recognized independently and statically using different classification models such as Neural Network [10, 19], Support Vector Machine [11, 1], AdaBoost, sparse representation based model [12], rule based model [15], etc. The common weakness of these methods is that they ignore the semantic relationships among the action units and therefore may generate inconsistent AU predictions. The second type of methods overcomes these limitations by modeling AU dependencies. Pantic *et al*. [16] propose to capture the AU relationships with a set of explicit rules and the AUs are recognized using a fast direct chaining inference procedure. Furthermore, temporal rules are also introduced in [15]. However, only a small number of rules are defined and the uncertainties of the rules are not well handled. Statistical methods are also proposed. Tong *et al*. [20] apply the Bayesian network (BN) to model the semantic relationships among the AUs. The dependencies among the AUs are discovered by learning the structure and the conditional probabilities of an AU network. Dynamic Bayesian network (DBN) is also proposed to further capture the temporal relationships among AUs in [21]. Nevertheless, due to the Markov assumption, both BN and DBN are limited to capture the local relationships between pairs of AUs such as co-occurrence, co-absence and mutual exclusion. Morover, it is difficult to obtain the optimal structure especially when working with a large number of AUs. Another issue of BN based models is that they fail to capture expression dependent AU dependencies. By drawing upon recent successful applications of the restricted Boltzmann machine, our model is able to capture not only local but also global AU dependencies. Furthermore we can also incorporate related factors to improve the understanding of the relationships among action units.

Another related work worth mentioning is proposed in [18], where a deep belief network is used to model the facial expressions and the latent units are used to capture higher level image representations. However, our model is essentially different from theirs since we use the latent units to model the dependencies among the action units.

3. Restricted Boltzmann Machine

The proposed algorithm is inspired by the nice property of RBM that it is able to model high-order dependencies. Before introducing the proposed algorithm, we briefly review RBM and its learning method.

RBM is composed of a layer of visible variables $\mathbf{v} \in \{0,1\}^n$ and a layer of hidden variables $\mathbf{h} \in \{0,1\}^m$. A graphical depiction of RBM is shown in Figure 2, where each latent node is connected to each visible node. RBM is essentially an undirected graphical model, of which the total energy is defined in Equation 1. $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ are the parameters. W_{ij} measures the compatibility between visible node v_i and latent node h_j. $\{b_i\}$ and $\{c_j\}$ are the biases of the visible and latent units respectively.

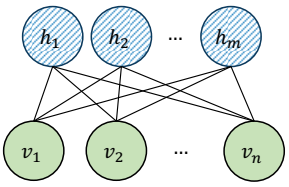

Figure 2: Restricted Boltzmann Machine

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_i \sum_j v_i W_{ij} h_j - \sum_i b_i v_i - \sum_j c_j h_j \quad (1)$$

Different from a Bayesian network which factorizes the joint distribution into the product of local probabilities, the distribution of the visible units of RBM is calculated by marginalizing over all the hidden units with Equation 2, where $Z(\theta)$ is the partition function. This allows to capture global dependencies among the visible variables instead of local relationships.

$$P(\mathbf{v}; \theta) = \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}|\theta))}{Z(\theta)} \quad (2)$$

Given the training data $\{\mathbf{v}_i\}_{i=1}^N$, the parameters are learned by maximizing the log likelihood with Equation 3.

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta); \quad \mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \log P(\mathbf{v}_i; \theta) \quad (3)$$

The gradient with respect to θ can calculated with Equation 4, where $\langle \cdot \rangle_p$ represents the expectation over distribution p.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h}|\mathbf{v},\theta)} - \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h},\mathbf{v}|\theta)} \quad (4)$$

Calculating the gradient involves inferring $P(\mathbf{h}, \mathbf{v})$ which is intractable. However it can be efficiently estimated with the contrastive divergence algorithm (CD) [5], and the basic idea to approximate $P(\mathbf{h}, \mathbf{v})$ with a one step Gibbs sampling from the data.

4. A Hierarchical Model for AU Recognition

RBM provides us with an effective tool to model high-order dependencies among the visible inputs. In this section we introduce our proposed algorithm. We first present our hierarchical model for AU recognition and then discuss each part of the model in detail.

Figure 3 shows the structure of the proposed model, which consists of three layers. The middle layer contains the binary visible units $\{a_1, \cdots, a_n\}$, representing the state of AU_1 to AU_n. A layer of binary latent units $\{h_1, \cdots, h_m\}$ are imposed upon the action units. Like an RBM, each latent unit is connected to all the AU nodes and therefore is used to model their global relationships. The variables $\{x_1, \cdots, x_d\}$ in the bottom layer stand for the image features. They are attached to each AU variable, providing low-level evidences.

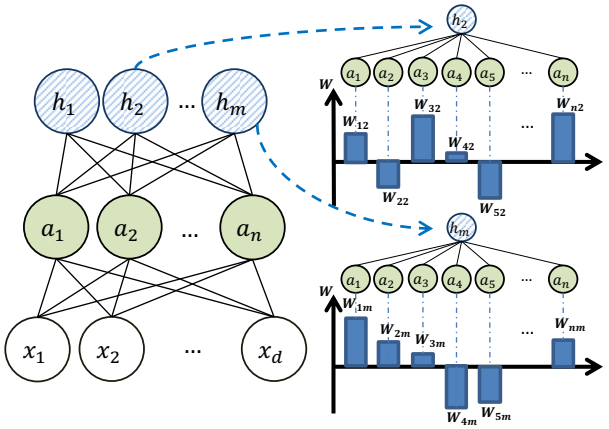

Figure 3: Proposed hierarchical model for joint facial action units recognition. Left: graphical depiction of the model. Right: the captured AU combination patterns of two latent units implied by their parameters.

The total energy of the model is defined in Equation 5, where $\{b_i\}$ and $\{c_j\}$ are the biases for the AU nodes and latent nodes respectively. The first set of parameters $\{W_{ij}^1\}$ measures the compatibility between each pair of latent and visible units (a_i, h_j), and the second set of parameters $\{W_{it}^2\}$ measures the compatibility between each pair of feature x_t and the i^{th} AU label a_i.

$$E(\mathbf{x}, \mathbf{a}, \mathbf{h}; \theta) = -\sum_i \sum_j a_i W_{ij}^1 h_j - \sum_j c_j h_j - \sum_i b_i a_i - \sum_i \sum_t W_{it}^2 a_i x_t \quad (5)$$

The proposed model can be decomposed into two parts with the first part consisting of the top two layers and the second part consisting the bottom two layers. In the following we discuss each part in detail.

Top down – capturing global relations among AUs. Unlike a regular BN, RBM is able to capture higher-order dependencies among the visible variables by connecting all the visible units through the latent units. Instantiating the visible units with the AU labels, we are able to capture the global relationships among the facial action units. Following this path, the top two layers of the proposed model assumes an RBM-like structure, with each latent unit connecting to all the action units to model their dependencies.

These two layers constitute the top part of the model and encode the high-level semantic relationships among AUs, enabling us to infer the presence or absence of each AU in the top-down direction.

The captured AU relationships can be implicitly inferred from the model parameters $\{\mathbf{W}_{ij}\}$ (bias terms are omitted without loss of generality). To gain some insight as to how they are related, consider the m^{th} latent unit h_m. Its compatibility with each action unit is measured by the pairwise energy $E(h_m, a_i) = -W_{im}a_i h_m$, $i = 1, \cdots, n$. We can see that larger W_{im} would lead to lower energy (thus higher probability). Therefore the larger W_{im} is, the more likely a_i will be present. Conversely, the smaller W_{im} is, the more likely a_i will be absent. As a whole, vector $[W_{im}]_{i=1}^n$ captures a specific presence and absence pattern of all the action units. Figure 3 graphically depicts the corresponding parameter vectors $[W_{im}]_{i=1}^n$ ($[W_{i2}]_{i=1}^n$) for h_m (h_2). In this case, h_m captures the pattern where a_1 is very likely to occur, $\{a_2, a_3, a_n\}$ tends to occur yet are less likely than a_1, and $\{a_4, a_5\}$ are very likely to be absent.

Bottom up – AU recognition from image features. The bottom two layers integrate the image shape or appearance features for AU recognition in the bottom-up direction. The features are attached to each AU node, and the relation between the features and the corresponding AU label is defined with the energy $E(a_i, \mathbf{x}) = -\sum_t W_{it}^2 a_i x_t$. If we remove the top layer, the second part is essentially equivalent to a set of linear AU classification models.

As a whole, the proposed model captures the information from both the top-down and bottom-up directions. Importantly, the captured relationships among the action units are global.

4.1. Model Learning

Learning of the proposed model amounts to parameter estimation. With the training data $\{(\mathbf{x}_i, \mathbf{a}_i)\}_{i=1}^N$, parameters are learned in a discriminative manner by maximizing the log conditional likelihood as shown in Equation 6. It is maximized with the stochastic gradient descent method in which the gradient can be calculated with Equation 7.

$$\theta^* = \arg\max_\theta \mathcal{L}(\theta); \mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^N \log P(\mathbf{a}_i|\mathbf{x}_i;\theta) \quad (6)$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h}|\mathbf{a},\mathbf{x},\theta)} - \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h},\mathbf{v}|\mathbf{x},\theta)} \quad (7)$$

Calculating the gradient involves inferring $P(\mathbf{h}|\mathbf{a},\mathbf{x},\theta)$ and $P(\mathbf{h},\mathbf{a}|\mathbf{x},\theta)$. $P(\mathbf{h}|\mathbf{a},\mathbf{x},\theta)$ can be analytically calculated with Equation 8, where $\sigma(x) = 1/(1+e^{-z})$ is the sigmoid function. $p(\mathbf{h},\mathbf{a}|\mathbf{x},\theta)$ is also intractable to compute. Therefore we extend the CD algorithm to learn the proposed model. The basic idea is to approximate $p(\mathbf{h},\mathbf{a}|\mathbf{x},\theta)$ by sampling $\mathbf{h}$ with Equation 8 and then sampling $\mathbf{a}$ with Equation 9. The detailed algorithm for learning the parameter $\mathbf{W}^1$ is shown in Algorithm 1. Other parameters can be estimated in the similar manner.

Algorithm 1 Revised contrastive divergence algorithm for learning the proposed model

Input: Training data $\{\mathbf{a}_i \in \mathbb{R}^{1\times n}, \mathbf{x}_i \in \mathbb{R}^{1\times d}\}_{i=1}^N$
Output: Model parameters $\mathbf{W}^1 \in \mathbb{R}^{n\times m}$
repeat
  Randomly pick a training instance $(\mathbf{a}, \mathbf{x})$
  Sample $\mathbf{h}^+ \sim P(\mathbf{h}|\mathbf{a},\mathbf{x})$ with Equation 8
  Calculate the positive gradient $D^+ = \mathbf{a}^T\mathbf{h}^+$
  Sample $\mathbf{a}^- \sim P(\mathbf{a}|\mathbf{h}^+,\mathbf{x})$ with Equation 9
  Sample $\mathbf{h}^- \sim P(\mathbf{h}|\mathbf{a}^-,\mathbf{x})$ with Equation 8
  Calculate the negative gradient $D^- = \mathbf{a}^{-T}\mathbf{h}^-$
  Update $\mathbf{W}^1 = \mathbf{W}^1 + \eta(D^+ - D^-)$
until Convergence

$$P(h_j|\mathbf{a},\mathbf{x}) = P(h_j|\mathbf{a}) = \sigma(-c_j - \sum_i W_{ij}^1 a_i) \quad (8)$$

$$P(a_i|\mathbf{h},\mathbf{x}) = \sigma(-b_i - \sum_j W_{ij}^1 h_j - \sum_t W_{it}^2 x_t) \quad (9)$$

Note that in this case the proposed model has the same formulation as a hidden conditional random field (HCRF), yet has a different structure from a regular HCRF. A regular HCRF impose an layer of latent units between the input and output to better model the intermediate feature structures. In our model however, the latent units are imposed in the top layer to capture the global relationships among the action units.

4.2. Inference

Given the query sample $\mathbf{x}$ during testing, we classify each action unit a_i by maximizing its posterior probability given $\mathbf{x}$ with Equation 10.

$$a_i^* = \arg\max_{a_i} P(a_i|\mathbf{x}) \quad (10)$$

Computing $P(a_i|\mathbf{x})$ requires marginalizing over all the latent variables $\{h_j\}_{j=1}^m$ and other action units $\{a_s\}_{s\neq i}$ which could be intractable. However it can be efficiently performed with the Gibbs sampling method by iteratively sampling $\mathbf{h}$ from $P(\mathbf{h}|\mathbf{a},\mathbf{x})$ and sampling $\mathbf{a}$ from $P(\mathbf{a}|\mathbf{h},\mathbf{x})$. Sampled instances of each a_i are used to calculate the corresponding marginal probability. Detailed steps are presented in Algorithm 2.

5. Incorporating Related Factors

The relationships among the action units depending on factors such as the expression, age and gender of the subject. These factors are usually widely available during training but not available during testing. Hence they are also known as privileged information [24]. In this section we

Algorithm 2 Inference of $P(\mathbf{a}|\mathbf{x})$ with Gibbs Sampling

Input: Test sample $\mathbf{x}$; Parameters $\mathbf{W}^1$, $\mathbf{W}^2$, $\mathbf{b}$, $\mathbf{c}$
Output: $P(a_i|\mathbf{x})$ for $i = 1, \cdots, n$
for $chain = 1 \rightarrow C$ **do**
  Randomly initialize $\mathbf{a}^0$
  for $t = 0 \rightarrow N$ **do**
    Sample $\mathbf{h}^t \sim P(\mathbf{h}|\mathbf{a}^t, \mathbf{x})$ with Equation 8
    Sample $\mathbf{a}^{t+1} \sim P(\mathbf{a}|\mathbf{h}^t, \mathbf{x})$ with Equation 9
  end for
end for
for $i = 1 \rightarrow n$ **do**
  Collect the last K samples of a_i from each chain
  Calculate $P(a_i|\mathbf{x})$ based on the collected samples
end for

demonstrate how we incorporate these factors during training to facilitate the estimation of the AU dependencies. Here we focus on the facial expressions, but the same approach is readily applicable to capture other related factors.

Denote the facial expression with a discrete variable $\mathbf{l} = [l_1, \cdots, l_K]$, with $l_k = 1$ representing the presence of the k^{th} expression. Inspired by the 3-way restricted Boltzmann machine [14], the basic idea is to modulate the connection between each pair of action unit and latent unit (a_i, h_j) with the expression variable $\mathbf{l}$, as shown in Figure 4a. Each clique in this case contains three variables (a_i, h_j, l_k), and their energy is defined as $E(a_i, h_j, l_k) = -W^1_{ijk} a_i h_j l_k$. We can see that the expression variable multiplicatively interacts with both the action unit variable and the latent variable to determine the energy of the model, and the parameters $\{W^1_{ijk}\}$ now form a 3D tensor instead of a matrix in the previous case. The graphical depiction of a more complex example is shown in Figure 4b, which contains two latent units, two action units and two expressions.

Since the image features also provide evidence about the expression, we also connect the feature variable $\mathbf{x}$ with the expression variable $\mathbf{l}$. The shorthand depiction of our proposed model to capture related factors is shown in Figure 4c. In the top level, the expression is incorporated to modulate our estimation of the AU relationships. In the bottom level, the image features provide information about both the type of expression and the AUs.

The total energy of the model is defined in Equation 11, with an additional term $-\sum_k \sum_t W^3_{kt} l_k x_t$ modeling the relationship between $\mathbf{x}$ and $\mathbf{l}$, and $-\sum_k d_k l_k$ modeling the bias of the expression node. Parameters are learned by maximizing the log conditional likelihood of both $\mathbf{a}$ and $\mathbf{l}$ with Equation 12.

Similarly, the CD algorithm can be extended to learn this model by iteratively sampling $\tilde{\mathbf{h}}$ from $P(\mathbf{h}|\mathbf{a}, \mathbf{l}, \mathbf{x})$ and sampling $(\tilde{\mathbf{a}}, \tilde{\mathbf{l}})$ from $P(\mathbf{a}, \mathbf{l}|\mathbf{h}, \mathbf{x})$. We follow the same procedure proposed in [14] to estimate the parameters. The

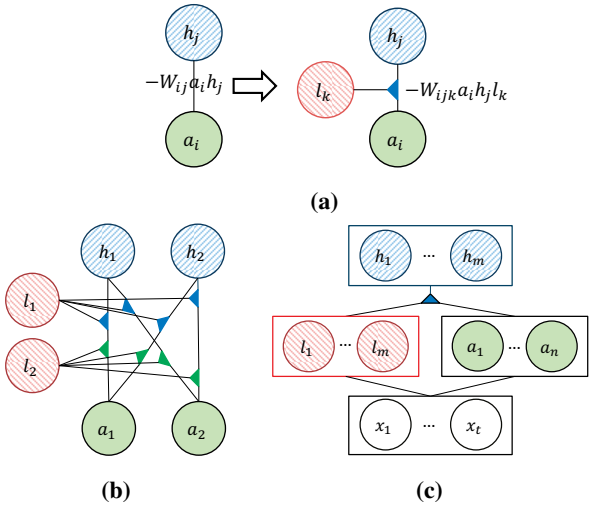

Figure 4: (a) Every clique in the proposed model contains a latent unit, an action unit and an expression unit. l_k multiplicatively modulate the connection between h_j and a_i. (b) An example with two action units, two latent units and two expressions. (c) Shorthand notation of the proposed model.

only revision we make is that during each step, we compute $P(\mathbf{h}|\mathbf{a}, \mathbf{l}, \mathbf{x})$ and $P(\mathbf{a}, \mathbf{l}|\mathbf{h}, \mathbf{x})$, instead of $P(\mathbf{h}|\mathbf{a}, \mathbf{l})$ and $P(\mathbf{a}, \mathbf{l}|\mathbf{h})$.

$$\begin{aligned} E(\mathbf{x}, \mathbf{a}, \mathbf{l}, \mathbf{h}; \theta) = &- \sum_i \sum_j \sum_k W^1_{ijk} a_i h_j l_k \\ &- \sum_i \sum_t W^2_{it} a_i x_i - \sum_k \sum_t W^3_{kt} l_k x_t \\ &- \sum_j c_j h_j - \sum_i b_i a_i - \sum_k d_k l_k \end{aligned} \tag{11}$$

$$\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log P(\mathbf{a}_i, \mathbf{l}_i|\mathbf{x}_i, \theta) \tag{12}$$

Given a query sample $\mathbf{x}$ during testing, each action unit a_i is also classified by maximizing its marginal posterior probability: $a_i^* = \arg\max_{a_i} P(a_i|\mathbf{x})$, which can still be efficiently calculated with the Gibbs sampling method.

6. Experiments

The goal of our experiments is to evaluate whether our proposed models can improve AU recognition over existing approaches by incorporating global AU semantics and related factors. For easy notation, we denote the proposed model described in Section 4 as **HRBM**, and the model described in Section 5 which incorporates expression information as **HRBM+**. We mainly compare our models with the baseline proposed by Tong *et al.* [20], which uses a Bayesian network to capture the AU relationships (**BN**). Since they only reported results on the Cohn-Kanade database (CK), results of BN on other data sets in our ex-

periments are reproduced with their provided code. We also compare with other related works. For efficiency, the raw features are first fed into a set of Support Vector Machines (**SVM**), each of which is trained independently to recognize one action unit. The output scores of these SVMs are then used as the input feature for other models. To make a thorough comparison, we test their performances on both posed and non-posed facial behavior databases.

6.1. Implementation Details

For the proposed models HRBM and HRBM+, parameters were randomly initialized in training. During our experiment we found that the recognition performance was not sensitive to the number of latent units. We chose 40 latent units in all our experiments. To infer the AU labels with Gibbs sampling, we used 5 Gibbs chains with each chain containing 10,000 steps. It took less than 0.2 seconds to infer all the AU labels for one instance on 2 cores of an Intel Core2 CPU E8400 @ 3.0GHz with 4 GB of memory.

6.2. Performance for Posed Facial Actions

First we evaluate the proposed models against the baselines on the extended Cohn-Kanade database (CK+) [11, 8], which contains 593 posed facial activity videos from 210 adults. Among all the participants, 9% are female, 81% are Euro-American, 13% are Afro-American and 6% are from other groups. All the peak frames are fully FACS coded. Seven expressions are labeled for 327 videos and they are happy, anger, surprise, fear, contempt, sad and disgust. All the rest sequences are treated as the eighth unknown expression in our experiment.

Following the procedure in [11], we extract both the appearance and shape features to recognize AUs of the peak frames in this database. Let $\{(x_i, y_i)\}_{i=1}^{c}$ and $\{(x_i^0, y_i^0)\}_{i=1}^{c}$ denote the facial feature points for a peak frame I and the corresponding neutral frame I_0, respectively. The shape feature for I is defined as $[x_1-x_1^0, y_1-y_1^0, \cdots, x_c-x_c^0, y_c-y_c^0]$. To extract the appearance feature, we first compute the mean shape of all the images and then align both image I and neutral image I_0 into the based shape through a patch-based piece-wise affine warping procedure [3]. We apply principal component analysis to their difference and select the components that remain at least 90% of the information. The coefficients of these principal components are used as the appearance feature.

The experiments are based on the leave-one-subject-out configuration. We use F_1-score to evaluate the performance of all models since it is a relatively fair measure for unbalanced data. The detailed results of different models for each action unit are illustrated in the left bar graph of Figure 5, and the average scores are shown in Table 1.

First, we can see that all the three models that capture AU relationships improve the performance of the baseline SVM for almost all the action units in different degrees. In particular, significant improvements are achieved for AUs that are difficult to recognize by individual classifiers (e.g. AU11, AU15, AU26). This demonstrates that the top-down information of AU relationships are especially useful when an AU is difficult to recognize from the measurements in the bottom-up direction. Second, by modeling higher-order AU interactions, HRBM further improves the performance of the BN-based method. Similarly, HRBM significantly outperforms BN for poorly recognized AUs. For instance, HRBM improves the F_1-score of BN from 36.51% to 49.09% for AU11, and 68.37% to 76.38% for AU15. Overall HRBM outperforms BN by about 2.5%. Finally, by properly incorporating the expression information only during training, HRBM+ further improves the recognition performance of HRBM by more than 3%. In total, the proposed algorithm improves the baseline SVM by about 7.5% and improves the performance of BN by about 6%.

Table 1: F_1-score of different models on CK+

Method	SVM	BN	HRBM	HRBM+
Average F_1-score	74.70%	76.70%	**79.21%**	**82.44%**

6.3. Performance for Non-Posed Facial Actions

To evaluate the generalization performance to real-world conditions, the next experiment is performed by applying the models trained on CK+ to the SEMAINE database [13]. Unlike CK+, the expressions of users in SEMAINE are naturally induced by operators during the conversation. Therefore the dataset contains speech related mouth and face movements, and significant amounts of both in- and out-of-plane head rotations. All these make the recognition task much more challenging. So far a total of 180 frames from 8 sessions of SEMAINE are FACS coded with experts, and in this experiment we recognize 10 AUs that are present for at least 15 times.

Following the work in [7], we use the Local Binary Pattern (LBP) feature in this part of experiment. The LBP feature is extracted in the same manner as [7]. The average F_1-scores of different models are shown in Table 2, and the F_1-scores for each individual AU are given in the right bar graph in Figure 5.

Table 2: F_1-score of different models on SEMAINE

Method	SVM	BN	HRBM	HRBM+
Average F_1-score	47.70%	51.09%	**54.76%**	**56.14%**

Similarly, while all the other three models improve the performance of the baseline SVM for all action units, the improvement achieved by HRBM is more than that of BN. In addition, by capturing the expression information HRBM+ further increases the average F_1-score of HRBM by about 2%. In total, the proposed method outperforms

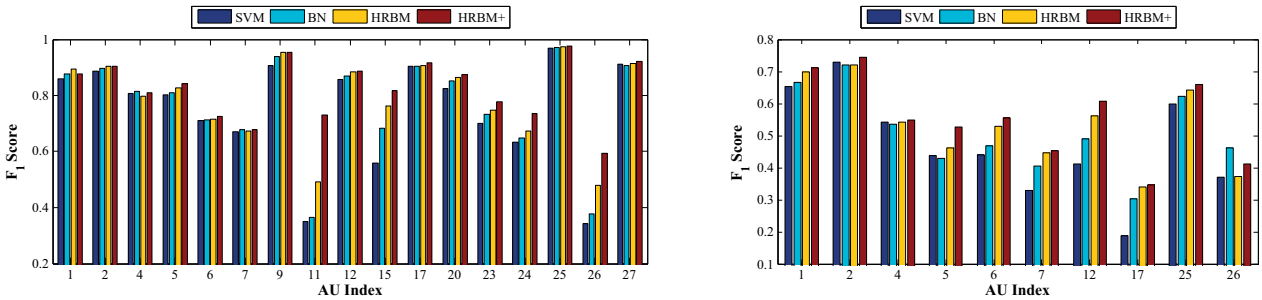

Figure 5: Comparison of different models for each action unit in terms of F_1-score. Left: results on CK+. Right: results on SEMAINE.

the baseline by about 8.5%. This demonstrates that the proposed algorithm can also more effectively benefit AU recognition than the BN-based method in real-world environments.

6.4. Comparison with Related Works

The proposed model outperforms BN for recognizing action units in both posed and non-posed facial behaviors. To gain some rough idea about the performance of our proposed model, we compare our method with some earlier works as listed in Table 3. On each dataset we use the same features, the same experimental configuration and the same evaluation criteria as the related works.

Several models were also used to capture AU relationships on CK dataset, including the Bayesian network (BN) and the dynamic Bayesian network (DBN). BN used in [20] captures the static relationships among AUs. DBN used in [21] further incorporates the temporal information of AUs. For comparison we reproduced the same features that were used by BN or DBN following [20] and applied HRBM for AU recognition. Since the expression is ambiguous between the neutral and peak frame, we did not implement HRBM+. From the results we can see that HRBM outperforms the performance of BN in [20]. Note that even without incorporating the dynamic information, our model can achieve slightly better performance than DBN.

Table 3: Comparison with Related Works

Author	Method	F_1	AUC	KSS
CK: image sequences				
Tong *et al.* [20]	BN	-	-	78.08%
Tong *et al.* [21]	DBN	-	-	79.71%
This work	HRBM	79.50%	-	**79.83%**
CK+: peak frames				
Lucey *et al.* [11]	SVM	-	94.5	-
This work	HRBM+	82.44%	**96.7**	-
SEMAINE: image frames				
Jiang *et al.* [7]	SVM	60.83%	-	-
This work	HRBM+	60.79%	-	-
CK+: peak frames				
Lucey *et al.* [11]	SVM	-	94.5	-
This work	HRBM+	82.44%	**96.7**	-
FERA: image frames				
Li *et al.* [9]	DBN	50.88%	-	-
This work	HRBM	**52.35%**	-	-

F_1 = F_1-score, AUC = area under the ROC curve, KSS = Hanssen -Kuiper Skill Score (true positive rate - false positive rate)

So far no works have been done to capture AU relationships on CK+ or SEMAINE. Instead we compare our model with some individual AU classifiers. On CK+, Lucey *et al.* [11] used SVM to perform AU recognition and achieved an average (AUC) area under the ROC curve of 94.5. In comparison, HRBM+ can achieve an AUC of 96.7. Using cross validation within the SEMAINE dataset, Jiang *et al.* [7] recognized the upper face AUs with the LBP features and reported an average F_1-score of 60.83%. Although HRBM+ is trained on CK+ and tested on SEMAINE, it can achieve a comparable F_1-score of 60.79% for the same upper face AUs. Finally, the proposed algorithm is also implemented on FERA [22], where we achieve a 2% improvement compared to a DBN model used by Li *et al.* [9]. All the above results demonstrate the competitive and promising performance of our proposed model on different datasets and different conditions, compared to other approaches.

6.5. Semantic Relationship Analysis

In addition to evaluating the AU recognition performance of the proposed model, we proceed to showing the captured semantic AU relationships. As discussed in Section 4, each latent unit captures a specific AU presence or absence pattern that is implied by the parameters $\{\mathbf{W}_{ij}\}$. Large W indicates high probability of occurrence and small W indicates high probability of absence. Parameters corresponding to two latent units of HRBM learned on the CK+ dataset are graphically illustrated in Figure 6. Unlike BN which can only encode pairwise AU dependencies, the proposed model captures higher-order presence or absence patterns that involve all the action units. For example in Figure 6, the first latent unit encodes the pattern that a person is likely to "lower brow", "wrinkle nose", "pull lip corner" and "drop jaw", but is unlikely to "depress lip corner" and "raise chin".

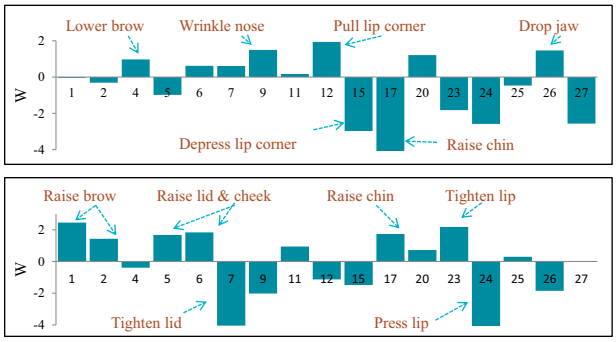

Figure 6: Two semantic AU relationship patterns captured by two latent units of HRBM. X-axis – AU index. Y-axis – value of parameter W. Large W indicates high probability of occurrence. Small W indicates high probability of absence.

7. Conclusion

In this paper we have proposed a hierarchical model which systematically integrates the low-level image measurements with the high-level AU semantical relationships for AU recognition. While existing methods can only capture local pairwise AU dependencies, the proposed model is built upon the restricted Boltzmann machine, and lends itself to capture higher-order AU interactions. The model is further developed to capture related factors such as the facial expressions to achieve better characterization of the AU relationships. Experimental results on both posed and non-posed facial action datasets demonstrated the power of the proposed model in capturing AU relationships as well as its advantage over existing methodologies for AU recognition. Moreover, the proposed methods are readily applicable to other applications that involve multiple related outputs.

Acknowledgement

This work is funded in part by NSF grant IIS 1145152, ARO grant W911NF-12-1-0473, and NSFC grant 61228304.

References

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *CVPR*, pages 568–573, 2005. 2

[2] J. Bazzo and M. Lamar. Recognizing facial actions using gabor wavelets with neutral face average difference. In *Automatic Face and Gesture Recognition, IEEE International Conference on*, pages 505–510, 2004. 2

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, Jun. 6

[4] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. 1

[5] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, Aug. 2002. 3

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006. 2

[7] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face Gesture Recognition and Workshops, IEEE International Conference on*, pages 314–321, 2011. 2, 6, 7

[8] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46 –53, 2000. 6

[9] Y. Li, Y. Zhao, S. Wang, and Q. Ji. Simultaneous facial feature tracking and facial expression recognition. *Image Processing, IEEE Transactions on*, 2013. 7

[10] J.-J. J. Lien, T. Kanade, J. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous Systems*, July 1999. 2

[11] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshop*, 2010. 2, 6, 7

[12] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *In Automatic Face & Gesture Recognition and Workshops, IEEE International Conference on*, pages 336–342. IEEE, 2011. 2

[13] G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder. The semaine database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, April 2012. Issue 1. 6

[14] V. Nair and G. E. Hinton. 3d object recognition with deep belief nets. In *NIPS*, pages 1339–1347, 2009. 2, 5

[15] M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):433–449, 2006. 2

[16] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1449 –1461, june 2004. 2

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, IEEE International Conference on*. IEEE, 2005.

[18] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson. Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440, 2008. 2

[19] Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001. 2

[20] Y. Tong and Q. Ji. Learning bayesian networks with qualitative constraints. In *CVPR*, pages 1–8. IEEE, 2008. 2, 5, 7

[21] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1683–1699, Oct. 2007. 2, 7

[22] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops, IEEE International Conference on*, pages 921–926. IEEE, 2011. 7

[23] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pages 28–43, 2012. 2

[24] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22:544–557, July 2009. 4

[25] J. Whitehill and C. Omlin. Haar features for facs au recognition. In *Automatic Face and Gesture Recognition, 7th International Conference on*, pages 97–101, 2006. 2