Facial Action Unit Recognition and Intensity Estimation Enhanced Through Label Dependencies

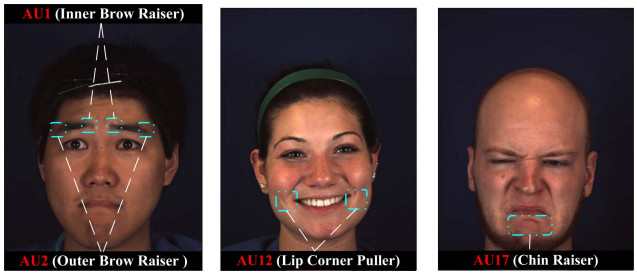
Shangfei Wang, *Senior Member, IEEE*, Longfei Hao, and Qiang Ji, *Fellow, IEEE*

***Abstract*— The inherent dependencies among facial action units (AUs) caused by the underlying anatomic mechanism are essential for the proper recognition of AUs and the estimation of intensity levels, but they have not been exploited to their full potential. We are proposing novel methods to recognize AUs and estimate intensity via hybrid Bayesian networks (BNs). The upper two layers are latent regression BNs (LRBNs), and the lower layers are BNs. The visible nodes of the LRBN layers are the representations of ground-truth AU occurrences or AU intensities. Through the directed connections from latent layer and visible layer, an LRBN can successfully represent relationships between multiple AUs or AU intensities. The lower layers include BNs with two nodes for AU recognition, and BNs with three nodes for AU intensity estimation. The bottom layers incorporate measurements from facial images with AU dependencies for intensity estimation and AU recognition. Efficient learning algorithms of the hybrid Bayesian networks are proposed for AU recognition as well as intensity estimation. Furthermore, the proposed hybrid BN models are extended for facial expression-assisted AU recognition and intensity estimation, as AU relationships are closely related to facial expressions. We test our methods on three benchmark databases for AU recognition and two benchmark databases for intensity estimation. The results demonstrate that the proposed approaches faithfully model the complex and global inherent AU dependencies, and the expression labels available only during training can boost the estimation of AU dependencies for both AU recognition and intensity estimation.**

***Index Terms*— AU recognition, AU intensity estimation, latent regression Bayesian network, label dependencies.**

I. Introduction

RECENT years have seen increasing research on automatic facial expression recognition and facial action unit (AU) analyses due to their great application potential in human-computer interaction. Facial expression categories are global descriptions of facial behavior. There has not previously been a complete, established expression category set. Facial action units describe facial behavior locally, representing the movement of one or more muscles in the face. An established, complete facial action unit set has already been proposed by Ekman and Friesen [1]. Almost all anatomically feasible facial expressions can be deconstructed into several AUs. This paper focuses on AU analyses.



(a) (b) (c)

Fig. 1. The examples of co-occurrent and mutual exclusive relations among AUs. (a) AU1+AU2. (b) AU12. (c) AU17.

Current work of AU analyses mainly detects action units or estimates the intensity of each action unit independently, ignoring AU relations. Since certain anatomic mechanisms govern the interactions between facial muscles, there are dependencies among the AUs, including both co-occurrence and mutual exclusion. Certain AU combinations may lead to certain facial expressions. For example, as shown in Fig. 1, both the inner brow raiser action unit (AU1) and the outer brow raiser action unit (AU2) are associated with the frontalis muscle group. Most people cannot move AU1 without also moving AU2, and vice versa. These action units have a coexistent relationship. On the other hand, the lip corner puller action unit (AU12) does not usually coincide with chin raiser (AU17). The former requires the zygomaticus major muscle group, while the latter uses the mentalis muscle group. They have a mutually exclusive relationship. Most people raise their cheeks and stretch their mouths when smiling. The Emotional Facial Action Coding System (EMFACS) has a list of AU combinations which often appear in certain expressions. These demonstrate the close relations between expressions and AUs. The inherent relationships between AUs themselves, as well as between expressions and AUs, should be exploited to improve AU recognition and AU intensity estimation.

Manuscript received February 18, 2018; revised September 10, 2018; accepted October 21, 2018. Date of publication October 26, 2018; date of current version November 21, 2018. This work has been supported by the National Science Foundation of China (Grant No. 61473270, 917418129, 61727809), and the project from Anhui Science and Technology Agency (1804a09020038). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Husrev T. Sencar. *(Corresponding author: Shangfei Wang).*

S. Wang is with the Key Laboratory of Computing and Communication Software of Anhui Province, the School of Computer Science and Technology, and the School of Data Science, University of Science and Technology of China, Hefei 230027, China (e-mail: sfwang@ustc.edu.cn).

L. Hao is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: hlf101@mai.ustc.edu.cn).

Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, NY 12180 USA (e-mail: qji@ecse.rpi.edu).

Digital Object Identifier 10.1109/TIP.2018.2878339

Researchers have just started exploring the role AU relations play in AU recognition and intensity estimation. Both discriminative approaches and generative approaches are investigated. The former incorporates further constraints on the loss function to represent AU relations, and the latter adopts the parameter and structure of probabilistic graphical models to capture the probabilistic dependencies between AUs. The additional constraints can model local or fixed AU relations, but are unable to model the many variations in AU dependencies. while probabilistic graphical models can capture more complex and global AU dependencies through their diversiform structures. Therefore, we prefer to adopt generative approaches for AU analysis enhanced by label dependencies.

Compared to currently used probabilistic graphical models like dynamic Bayesian networks (DBNs) [2], [3] and restricted Boltzmann machines (RBMs) [4]–[6], latent regression Bayesian networks (LRBNs) [7] can more thoroughly represent visible units. The LRBN is able to take into account the global dependencies between visible nodes, as well as the dependencies among hidden nodes given observations. In this paper, we capture high-order and global dependencies among AUs by using an LRBN. Specifically, we propose a hybrid Bayesian network in which the upper two layers consist of latent regression Bayesian network and the lower layers are composed of Bayesian networks. The ground-truth AU states or AU intensities are represented by the visible nodes of the LRBN. Through the learning process, the LRBN is able to accurately represent relationships among multiple AUs or AU intensities. The lower layers include two-node Bayesian networks for AU recognition, and three-node Bayesian networks for AU intensity estimation. The bottom layers perform AU recognition and intensity estimation by incorporating measurements from facial images with AU dependencies. We then extend our proposed hybrid BN model, using the AU relations and expression-AU relations for AU recognition and intensity estimation. By introducing facial expression nodes as visible variables, facial expressions, only available during training, can more effectively capture AU dependencies. Experimental results on several benchmark databases show that our proposed approaches are able to effectively capture complicated intrinsic relationships among AUs. We further show that expression labels can more effectively model AU dependencies for AU recognition and intensity estimation, even when the labels are only available during training.

The rest of this paper is organized in the following manner. The next section gives an overview of the related work on AU recognition and AU intensity estimation. Section III briefly introduces LRBN. Section IV and Section V elaborate on the proposed AU recognition models and AU intensity estimation models, respectively. Section VI presents the experimental results on three databases for AU recognition and two databases for intensity estimation, and makes the comparison to related works. Section VII concludes our work.

II. Related Work

A recent comprehensive survey of facial action unit analysis can be found in Martinez *et al.* [8]. In this section, we briefly review AU recognition and intensity estimation works that leverage label dependencies.

A. AU Recognition Leveraging Label Dependencies

Researchers have recently begun to examine the ways in which AU relations could improve AU recognition. Both discriminative approaches and generative approaches are investigated. For discriminative approaches, additional constrains of the loss function are used to represent AU relations. Zhu *et al.* [9] and Zhang and Mahoor [10] regarded each AU recognition as a singular task, and adopted multi-task learning for the simultaneous recognition of multiple AUs. The constraints among several tasks were representative of fixed and local AU relationships. Zhao *et al.* [11] employed the constraints to represent group sparsity as well as local positive correlation and negative competition for multiple AU recognition. Eleftheriadis *et al.* [12] suggested a multi-conditional latent variable model. This model projects the features of the image onto a shared manifold, which is then regularized by constraints representing global and local co-occurring dependencies among the AU labels. Eleftheriadis *et al.* considered co-occurrent relationships between AU labels, without considering mutually exclusive relationships. All the adopted constrains can model certain kind of AU dependencies, but cannot fully represent hundreds of variations in AU dependencies.

Generative approaches use the parameters and structure of probabilistic graphical models to incorporate the probabilistic dependencies between AUs. Tong *et al.* [2] and Li *et al.* [3] each proposed a dynamic Bayesian network (DBN) to capture probabilistic relationships and temporal changes among action units. Because of the Markov assumption, their proposed DBNs are only able to capture local relations, such as co-occurrence and mutual exclusion, between pairs of AUs. Wang *et al.* [4] and Wu *et al.* [5], [6] proposed an RBM to capture the global relations among AUs, since RBM introduces a layer of latent units in order to model higher-order dependencies among random variables. However, as an undirected latent variable model, hidden units of RBM are independent to one other given the visible units. The introduction of dependencies among the hidden units is expected to allow the model to better explain the patterns embedded in the visible units. Unlike an RBM, an LRBN is a directed model, capturing the global dependencies among visible nodes as well as the dependencies among hidden nodes given observations. It therefore offers a better representation of visible units via directed links amongst hidden and visible units. Therefore, we employ an LRBN to successfully model high-order and global AU dependencies.

Unlike AU recognition enhanced by AU relations, expression-assisted AU recognition has not been paid as much attention. As far as we know, just three works recognize action units assisted by expressions. One adopted discriminative approaches, and the other two used generative approaches. Ruiz *et al.* [13] proposed a discriminative approach to learn AU classifiers from unannotated facial images and another large-scale facial images with expression labels only. Their approach uses ground-truth expression labels to generate pseudo-labels according to summarized dependencies

between expressions and AUs from Gosselin *et al.* [14] and Scherer *et al.* [15]. Then, expression classifiers from AUs are trained with the generated AU pseudo-labels and the ground-truth labels. After that, AU classifiers are learned by using the output of the action unit classifiers as the input for the expression classifiers. Their research exploited the fixed and pairwise expression relations to train AU classifiers from facial images without AU annotations.

For generative approaches, a Bayesian network [16] and a three-way Restricted Boltzmann Machine [4] are used to capture the dependencies between AU and expressions in addition to the dependencies between AUs. Wang *et al.* [4], [17] proposed a Bayesian network to capture local dependencies between AUs and between AUs and expressions, for the task of AU recognition for images with full expression labels but limited action unit labels. The dependencies between expressions and AUs are used as supplementary to missed AU labels. Wang *et al.* [4] proposed a mixture model — a three-way RBM that independently captured the relationships between each expression and the AUs for AU recognition. The expression labels are only needed during training as privileged information.

The adopted Bayesian network captures pairwise relations between expressions and AUs, and the used three-way RBM models global relations among each expression and AUs. However, AU recognition requires the capture of more complex, global dependencies among expressions and AUs. Therefore, we extend the proposed LRBN to enable the capture of high-order and global AU dependencies and AU-expression dependencies for the task of AU recognition. Expression labels are only required during training.

B. AU Intensity Estimation Leveraging Label Dependencies

Due to the limited available databases with AU intensity annotations, few works consider AU intensity estimation. Among them, several works focus on the use of AU relations for intensity estimation using either discriminative approaches or generative approaches.

As in AU recognition, the additional loss function constraints as well as parameters and structures of generative models are used to capture dependencies among action units for intensity estimation. For example, Nicolle *et al.* [18] and Wang *et al.* [19] considered intensity estimation of one AU as one task, and proposed multitask learning solutions to predict AU intensities. The constrains among multiple tasks represent the local and fixed AU relations among a task group. Li *et al.* [20] used DBN to capture local and pairwise AU relationships in order to measure their intensities. Sandbach *et al.* [21] constructed Markov random field trees to depict the pairwise AU intensity combinations in the region of the upper face. Kaltwang *et al.* [22] proposed a generative latent tree model to represent the joint distribution of AU intensities and facial features to estimate multiple AU intensities. Walecki *et al.* [23] proposed a conditional random field (CRF) to model individual independent AUs as well as pairs of AUs to estimate intensity. Rudovic *et al.* [24] proposed a conditional ordinal random field model for context-sensitive modeling of AU intensity. Although current works explore AU dependencies to some extend for AU intensity estimation, more complete and global dependencies are still expected to explore. Therefore, we propose to employ the LRBN to capture and leverage high-order and global dependencies among AUs for improved AU intensity estimation.

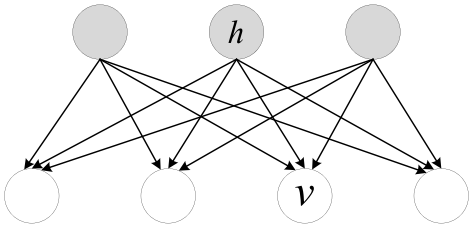


Fig. 2. The structure of LRBN.

As far as we know, just one work explores AU-expression dependencies for AU intensity estimation. Wang *et al.* [17] extended their method for expression-assisted AU recognition to expression-assisted AU intensity estimation. They adopted a Bayesian network to capture the local and pairwise dependencies between AU intensities and expressions. A structured EM is used to learn the parameters and structure of the Bayesian network for missing AU intensities. Expression labels are used to handle incomplete AU intensity labeling. Their work can only capture local and pairwise AU-expression dependencies. More global and complex AU-expression dependencies are expected to beneficial for AU intensity estimation. Therefore, we extend the proposed LRBN to capture both high-order and global dependencies among AUs, as well as among expressions and AUs. Expression labels are only required during training.

A previous version of the paper appeared as Hao *et al.* [25], which proposed a hybrid Bayesian network consisting of a LRBN and two-node BNs for AU recognition enhanced by AU dependencies and AU-expression dependencies. Compared with the previous version, this paper proposes another hybrid Bayesian network consisting of a LRBN and three-node BNs for AU intensity estimation through leveraging dependencies among AUs as well as among AUs and expressions. To show the effectiveness of the proposed intensity estimation method, experiments are conducted on the BP4D and PAIN databases.

III. Brief Introduction to LRBN

A latent regression Bayesian network is a type of directed latent graphic model. An LRBN is made up of a visible layer, a latent layer, and directed edges between visible nodes hidden nodes, as shown in Fig. 2. Since the "explaining away" effect can reduce the necessity of invoking alternative causes when one cause of an observed event is confirmed [26], the latent variables are independent of each other given the visible variables.

Bayesian networks are subject to the chain rule, wherein the joint probability of all visible and latent variables of a LRBN can be factorized into the product of prior probabilities for any latent variable h_j, $P(h_j)$, and the conditional probabilities of any visible node v_i given all latent variables $\boldsymbol{h}$, $P(v_i|\boldsymbol{h})$ as

shown in Eq. (1):

$$P(\boldsymbol{v},\boldsymbol{h}) = \prod_{j=1}^{n_h} P(h_j) \prod_{i=1}^{n_v} P(v_i|\boldsymbol{h}), \tag{1}$$

where n_h, n_v represent the number of hidden and visible nodes, respectively.

In our work, since both h and v are binary, $P(h_j)$ and $P(v_i|\boldsymbol{h})$ are assumed to the Bernoulli distribution. They can respectively be written as Eq. (2) and Eq. (3).

$$P(h_j) = \sigma(d_j)^{h_j}(1-\sigma(d_j))^{1-h_j}, \tag{2}$$

where $\sigma(x) = 1/(1+exp(-x))$, and d_j is the bias of the variable h_j.

$$P(v_i|\boldsymbol{h}) = \sigma(w_i^T\boldsymbol{h}+b_i)^{v_i}(1-\sigma(w_i^T\boldsymbol{h}+b_i))^{1-v_i}, \tag{3}$$

where w_i represents the weight between all of the latent nodes h and the visible node v_i, and b_i is the bias term for v_i.

Integrating Eq. (2) and Eq. (3) into Eq. (1) results in Eq. (4):

$$\begin{aligned} P_{\Theta_{LRBN}}(\boldsymbol{v},\boldsymbol{h}) &= \prod_j \frac{\exp(d_j h_j)}{1+\exp(d_j)} \prod_i \frac{\exp((w_i^T\boldsymbol{h}+b_i)v_i)}{1+\exp(w_i^T\boldsymbol{h}+\boldsymbol{b})} \\ &= \frac{\exp(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{v},\boldsymbol{h}))}{\prod_j\left(1+\exp(d_j)\right)}, \end{aligned} \tag{4}$$

where $\Theta_{LRBN} = \{\boldsymbol{W},\boldsymbol{b},\boldsymbol{d}\}$, and

$$\begin{aligned} \Gamma_{\Theta_{LRBN}}(\boldsymbol{v},\boldsymbol{h}) = &-\sum_i \left(w_i^T\boldsymbol{h}+b_i\right)v_i - \sum_j d_j h_j \\ &+\sum_i log(1+exp(w_i^T\boldsymbol{h}+b_i)). \end{aligned} \tag{5}$$

Compared to the energy function of an RBM, Eq. (5) has the extra term $\sum_i log(1+exp(w_i^T\boldsymbol{h}+b_i))$ which is used to explicitly capture relationships among latent variables. Unlike the RBM, which uses undirected links between visible and hidden nodes, the LRBN uses directed links. These directed links result in the dependencies among the latent layer given the visible layer. Thus, the LRBN is better able to explain the patterns inherent in the observations. Additionally, the LRBN does not suffer from an intractable partition function issue; the joint distribution is obtained using the product of all the prior and conditional probabilities.

IV. AU Recognition Enhanced via Label Dependencies

Fig. 3 shows the proposed hybrid BN for AU recognition enhanced by modeling AU relationships. The proposed network has three layers. The upper two layers is a latent regression Bayesian network, and the lower two layers are two-node Bayesian networks. To be specific, the visible nodes (i.e., $\boldsymbol{a}$), are representative of the ground-truth AU states. LRBN is able to capture the dependencies among latent variables and among the visible variables via the directed connections from the hidden nodes to the visible nodes. Thus, it is able to thoroughly capture the global and complex relations among several AUs. The lower two layers consist of two-node Bayesian networks. They connect the ground-truth AU states ($\boldsymbol{a}$), and their measurements ($\boldsymbol{a}'$). Therefore, the $\boldsymbol{a}$ and $\boldsymbol{a}'$ are binary. Using measurements obtained from the images as evidence, the lowest two-layer Bayesian network integrates the facial images with the AU dependencies for improved AU recognition.

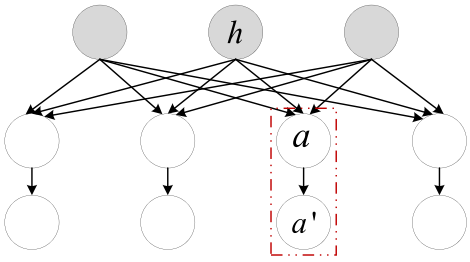


Fig. 3. Proposed AU recognition using AU-relation modeling.

The joint probability of all of the variables for the proposed hybrid Bayesian network is shown in Eq. (6):

$$\begin{aligned} P(\boldsymbol{a},\boldsymbol{h},\boldsymbol{a}') &= P(\boldsymbol{a}'|\boldsymbol{a})P(\boldsymbol{a},\boldsymbol{h}) \\ &= P(\boldsymbol{a}'|\boldsymbol{a})\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{a},\boldsymbol{h})}{\Pi_j(1+\exp(d_j))}. \end{aligned} \tag{6}$$

A. Parameter Learning

Given a set of samples $\mathcal{D} = \{\boldsymbol{a}^{(m)},\boldsymbol{a}'^{(m)}\}_{m=1}^M$, the parameters Θ are estimated by maximizing the marginal log-likelihood during training, according to Eq. (7):

$$\begin{aligned} \Theta^\star &= \arg\max_\Theta \frac{1}{M}\sum_{m=1}^M \mathcal{L}(\Theta) \\ &= \arg\max_\Theta \frac{1}{M}\sum_{m=1}^M \log(\sum_{\boldsymbol{h}} P(\boldsymbol{a},\boldsymbol{h},\boldsymbol{a}')) \\ &= \arg\max_\Theta \frac{1}{M}\sum_{m=1}^M \log(\sum_{\boldsymbol{h}} P(\boldsymbol{a}'|\boldsymbol{a})P(\boldsymbol{a},\boldsymbol{h})) \\ &= \arg\max_\Theta \frac{1}{M}\sum_{m=1}^M (\log P(\boldsymbol{a}'|\boldsymbol{a}) + \log\sum_{\boldsymbol{h}} P(\boldsymbol{a},\boldsymbol{h})). \end{aligned} \tag{7}$$

Eq. (7) demonstrates that the parameters of the uppermost two-layer (Θ_{LRBN}) and the parameters of the bottom two-layer BNs (Θ_{BN}) can be separately learned (i.e., $\Theta = \{\Theta_{BN},\Theta_{LRBN}\}$) as shown in Eq. (8) and Eq. (9),

$$\Theta^\star_{LRBN} = \arg\max_{\Theta_{LRBN}} \frac{1}{M}\sum_{m=1}^M \log(\sum_{\boldsymbol{h}} \frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{a},\boldsymbol{h})}{\Pi_j(1+\exp(d_j))}), \tag{8}$$

$$\Theta^\star_{BN} = \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^M \log P(\boldsymbol{a}'|\boldsymbol{a}). \tag{9}$$

1) Parameter Learning for the Uppermost Two-Layer LRBN: The gradient of Eq. (8) with respect to parameter Θ_{LRBN} is shown in Eq. (10),

$$\nabla_{\Theta_{LRBN}}\mathcal{L}(\Theta_{LRBN}) = \sum_m\sum_h P(\boldsymbol{h}|\boldsymbol{a}^{(m)})\frac{\partial -\Gamma_{\Theta_{LRBN}}(\boldsymbol{a}^{(m)},\boldsymbol{h})}{\partial\Theta_{LRBN}}. \tag{10}$$

The directed connections of LRBN are from hidden nodes to visible nodes, therefore, $P(\boldsymbol{h}|\boldsymbol{a}^{(m)})$ is computationally intractable. The exact gradient in Eq. (10) also requires exponential summations over all of the possible latent variables $\boldsymbol{h}$.

We prefer to obtain $P(\boldsymbol{h}|\boldsymbol{a}^{(m)})$ by adopting sampling method from the true posterior probability. This allows us to preserve certain dependencies among hidden variables. As the specific form of $P(h|a)$ is unavailable, it would be intractable to draw exact samples from $P(\boldsymbol{h}|\boldsymbol{a})$ through Gibbs sampling. Therefore, some approximations are taken during sampling according to Eq. (11).

$$\begin{aligned} P(\boldsymbol{h}|\boldsymbol{a}) &= \prod_j P(h_j|h1, \ldots, h_{j-1}, \boldsymbol{a}) \\ &\approx \prod_j P(h_j|h_{-j}, \boldsymbol{a}), \end{aligned} \tag{11}$$

where $h_{-j} = \{h_1, \ldots, h_{j-1}, h_{j+1}, \ldots, h_{n_h}\}$ is a set of all latent variables with the exception of h_j. Each of the latent nodes is sampled with all other nodes fixed as Eq. (12). Therefore, we are able to preserve the dependencies among latent variables to some extent. The procedure is iterated until convergence, whereupon a sample is collected and used to update the parameters.

$$h_j^t \sim P(h_j|\boldsymbol{a}, \boldsymbol{h}_{-j}^{t-1}). \tag{12}$$

Markov Chain Monte Carlo (MCMC) methods are typically used to estimate the summation with samples, thus addressing the issue of exponential summation. An intuitive estimation is shown in Eq. (13),

$$\nabla\Theta_{LRBN}\mathcal{L}(\Theta_{LRBN}) \approx \frac{1}{n}\sum_m\sum_s \frac{\partial - \Gamma_{\Theta_{LRBN}}(\boldsymbol{a}^{(m)}, \boldsymbol{h}^{(m,s)})}{\partial\Theta_{LRBN}}, \tag{13}$$

where $\boldsymbol{h}^{m,1}, \ldots, \boldsymbol{h}^{m,n}$ are n samples from $P(\boldsymbol{h}|\boldsymbol{a}^{(m)})$.

To avoid the computing complex of multiple Gibbs chains, we adopt the stochastic approximation procedure (SAP) framework [27]. The SAP only requires one latent variable sample for gradient estimation, and convergence to a local optimum is guaranteed [28] if the learning rate η_t satisfies Eq. (14),

$$\begin{aligned} \sum_{t=1}^{\infty}\eta_t &= \infty, \\ \sum_{t=1}^{\infty}\eta_t^2 &< \infty. \end{aligned} \tag{14}$$

The stochastic gradient ascent algorithm speeds up the learning phase, and we use a mini-batch of training samples to estimate the gradient.

The detailed algorithm for learning parameters Θ_{LRBN} can be seen in Algorithm 1.

2) Parameter Learning for the Lower Two-Layer BN: Due to the independence among different AU state nodes at the lower two-layer BN, Eq. (9) is rewritten as Eq. (15),

Algorithm 1 Parameter Learning for an LRBN [29]

Input training data $\mathscr{D} = \{\boldsymbol{a}^{(m)}\}_{m=1}^{M}$;
Output parameters Θ_{LRBN}.
1: Randomly initialize the parameters $\Theta_{LRBN} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{d}\}$;
2: Generate Gibbs samples at time step 0;
3: **while** parameters not converged, **do**
4: Randomly choose a batch of data samples $(\boldsymbol{a})$;
5: Perform Gibbs sampling to obtain one sample of the latent variables for one input data, $\boldsymbol{h}^{(t)} \sim P(\boldsymbol{h}|\boldsymbol{a}, \boldsymbol{h}^{(t-1)})$;
6: Compute the gradient;
7: Update the parameters, $\Theta_t = \Theta_{t-1} + \eta_t \bigtriangledown_{\Theta_{LRBN}} \mathscr{L}(\Theta_{LRBN})$.
8: **end while**
9: **return** Θ_{LRBN}

$$\begin{aligned} \Theta_{BN}^{\star} &= \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M} log\, P(\boldsymbol{a}'|\boldsymbol{a}) \\ &= \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{n}\log P(a_1', \ldots, a_n'|a_1, \ldots, a_n; \Theta_{BN}) \\ &= \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{n}\sum_{i=1}^{n}\log P(a_i'|a_i; \Theta_{BN})), \end{aligned} \tag{15}$$

where $a_1, a_2, \ldots, a_n$ are n ground-truth AU states, and $a_1', a_2', \ldots, a_n'$ are the corresponding measurements. All of them are available during training. Thus, maximum log-likelihood is used to obtain parameter Θ_{BN}.

B. Network Inference

The learned hybrid BN model combines AU measurements with the dependencies captured during training to recognize multiple AUs through inference. Specifically, we compute all possible AU combinations according to $P(\boldsymbol{a}|\boldsymbol{a}', \Theta)$, and then select the most probable explanation AU combination (i.e., the most probable explanation).

$$\begin{aligned} &a_1, \ldots, a_n \\ &= \arg\max_{a_1,\ldots,a_n} p(a_1, \ldots, a_n|a_1', \ldots, a_n') \\ &= \arg\max_{a_1,\ldots,a_n} \frac{p(a_1', \ldots, a_n'|a_1, \ldots, a_n)p(a_1, \ldots, a_n)}{p(a_1', \ldots, a_n')} \\ &= \arg\max_{a_1,\ldots,a_n} p(a_1', \ldots, a_n'|a_1, \ldots, a_n)p(a_1, \ldots, a_n) \\ &= \arg\max_{a_1,\ldots,a_n} \prod_{i=1}^{n} p(a_i'|a_i) \sum_{h_1,\ldots,h_m} p(a_1, \ldots, a_n, h_1, \ldots, h_m) \\ &= \arg\max_{a_1,\ldots,a_n} \prod_{i=1}^{n} p(a_i'|a_i) \\ &\quad \times \frac{\exp(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{a}, \boldsymbol{h}))}{\prod_i \left(1 + \exp(w_i^T + b)\right)\prod_j \left(1 + \exp(d_j)\right)}, \end{aligned} \tag{16}$$

where, $h_1, \ldots, h_m$ is m latent nodes of LRBN.

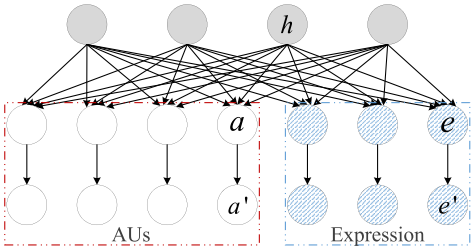


Fig. 4. The proposed AU recognition model enhanced by expression.

C. Expression-Assisted AU Recognition

As discussed in Section I, AU dependencies are influenced by expressions, since certain AU combinations formulate certain expressions. Such expression dependent AU relations can help AU recognition. Therefore, we extend our proposed AU recognition method, which is augmented by AU dependencies, to expression-assisted AU recognition method, which is enhanced by dependencies among AUs as well as dependencies among expressions and AUs as shown in Fig. 4.

Compared with Fig. 3, in which the second layer consists entirely of AU nodes, the second layer of the expression-assisted AU recognition model includes AU nodes $\boldsymbol{a}$ and expression nodes $\boldsymbol{e}$. Accordingly, the lowest layer contains AU measurement $\boldsymbol{a'}$ as well as expression measurement $\boldsymbol{e'}$. The upper two-layer LRBN thoroughly models global relationships among multiple AUs and among AUs and expressions via the directed connections from the hidden nodes to the expression and AU nodes. The lower two-layer network employs measurements taken from the images as evidence to integrate the facial images with the AU dependencies and the AU-expression relationships.

For our proposed expression-assisted AU recognition model, the joint probability of all is shown in Eq. (17):

$$\begin{aligned} &P(\boldsymbol{a},\boldsymbol{e},\boldsymbol{h},\boldsymbol{a'},\boldsymbol{e'}) \\ &\quad = P(\boldsymbol{a'},\boldsymbol{e'}|\boldsymbol{a},\boldsymbol{e})P(\boldsymbol{a},\boldsymbol{e},\boldsymbol{h}) \\ &\quad = P(\boldsymbol{a'},\boldsymbol{e'}|\boldsymbol{a},\boldsymbol{e})\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{a},\boldsymbol{e},\boldsymbol{h})}{\Pi_j(1+\exp(d_j))}. \end{aligned} \tag{17}$$

For both the AU recognition model and the proposed expression-assisted model, the parameters of the upper two-layer LRBN and those of the bottom two-layer Bayesian networks can be learned separately according to Eq. (18) and Eq. (19).

$$\Theta^{\star}_{LRBN} = \arg\max_{\Theta_{LRBN}} \frac{1}{M}\sum_{m=1}^{M}\log(\sum_{\boldsymbol{h}}\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{a},\boldsymbol{e},\boldsymbol{h})}{\Pi_j(1+\exp(d_j))}). \tag{18}$$

$$\begin{aligned} \Theta^{\star}_{BN} = \max_{\Theta_{BN}}\arg(&\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}\log P(a'_i|a_i;\Theta_{BN}) \\ &+\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_e}\sum_{j=1}^{n_e}\log P(e'_j|e_j;\Theta_{BN})), \end{aligned} \tag{19}$$

where n_e is the number of expression nodes $\boldsymbol{e}$.

Algorithm 2 Parameter Learning for an LRBN With Both AU Nodes and Expression Nodes

Input Training data $\mathscr{D} = \{\boldsymbol{a}^m, \boldsymbol{e}^m\}_{m=1}^{M}$;
Output Parameters Θ .

1: Randomly initialize the parameters $\Theta_{LRBN} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{d}\}$;
2: Generate Gibbs samples at the time step 0;
3: **while** parameters not converged, **do**
4: Randomly choose a batch of data samples $(\boldsymbol{a}, \boldsymbol{e})$;
5: Perform Gibbs sampling to obtain one sample of the latent variables for one input data, $\boldsymbol{h}^{(t)} \sim P(\boldsymbol{h}|\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{h}^{(t-1)})$;
6: Compute the gradient;
7: Update the parameters, $\Theta_t = \Theta_{t-1} + \eta_t \bigtriangledown_{\Theta_{LRBN}} \mathscr{L}(\Theta_{LRBN})$.
8: **end while**
9: **return** Θ_{LRBN}

For Eq. (18), similar to Algorithm 1, Algorithm 2 is proposed to learn the parameters Θ_{LRBN}. For Eq. (19), maximum log-likelihood is used to obtain parameter Θ_{BN}.

After the expression-assisted model is learned, AU measurements and expression measurements are combined with the captured AU dependencies and AU-expression dependencies to perform multiple AU recognition using probabilistic inference, according to Eq. (20):

$$\begin{aligned} &a_1,\ldots,a_n \\ &\quad = \arg\max_{a_1,\ldots,a_n}\prod_{i=1}^{n}P(a'_i|a_i) \\ &\qquad \times \max_{\boldsymbol{e}}\{\prod_{j=1}^{n_e}P(e'_j|e_j)\times\sum_{h_1..h_m}p(a_1..a_n,e_1..e_{n_e},h_1..h_m)\} \\ &\quad = \arg\max_{a_1,\ldots,a_n}\prod_{i=1}^{n}P(a'_i|a_i)\times\max_{\boldsymbol{e}}\{\prod_{j=1}^{n_e}P(e'_j|e_j) \\ &\qquad \times\frac{\exp(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{a},\boldsymbol{e},\boldsymbol{h}))}{\prod_i\left(1+\exp(w_i^T+b)\right)\prod_j\left(1+\exp(d_j)\right)}\}. \end{aligned} \tag{20}$$

Comparing the proposed AU recognition model with the expression-assisted AU recognition model, we see that the proposed AU recognition through AU-relation modeling happens to be a particular case of our proposed expression-assisted AU-recognition model, since expression nodes are excluded from the second and third layers.

V. AU Intensity Estimation via AU-Relation Modeling

Compared to action unit occurrences, AU intensities provide more fine-grained level for facial analysis. In this section, we propose AU intensity estimation model as show in Fig. 5. For simplicity, the top two layers are the same as those in Fig. 3. It means the visible nodes of LRBN in Fig. 5 are binary value. We use $\boldsymbol{v_a}$ to represent whether the ground-truth AU intensity is larger than its mean value or not. We further propose to a three-nodes Bayesian Network (tri-BN) [29] to capture the relations between ground-truth AU intensity and its measurement for AU intensity estimation. As shown in Fig. 5,

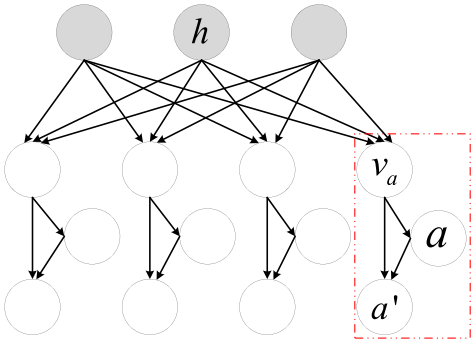


Fig. 5. The proposed estimation of AU intensity using AU-relation modeling.

the tri-BN consists of three nodes: one two-value node v_a and two multi-value discrete nodes a, a'. The node a represents the ground-truth AU intensity, and $\boldsymbol{a'}$ are its corresponding measurement.

The joint probability of all variables for Fig. 5 is as Eq. (21):

$$\begin{aligned} &P(\boldsymbol{v_a}, \boldsymbol{a}, \boldsymbol{a'}, \boldsymbol{h}) \\ &= P(\boldsymbol{a}|\boldsymbol{v_a})P(\boldsymbol{a'}|\boldsymbol{v_a}, \boldsymbol{a})P(\boldsymbol{v_a}, \boldsymbol{h}) \\ &= P(\boldsymbol{a}|\boldsymbol{v_a})P(\boldsymbol{a'}|\boldsymbol{v}, \boldsymbol{a})\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{h})}{\prod_j(1+\exp(d_j))}. \end{aligned} \tag{21}$$

A. Parameter Learning

The parameters of the upper two-layer LRBN and the lower tri-BN can be learned through Eq. (22) and Eq. (23), respectively.

$$\Theta^*_{LRBN} = \arg\max_{\Theta_{LRBN}} \frac{1}{M}\sum_{m=1}^{M}\log(\sum_{\boldsymbol{h}}\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{h})}{\prod_j(1+\exp(d_j))}), \tag{22}$$

$$\begin{aligned} \Theta^*_{BN} &= \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M}\log P(\boldsymbol{v_a})P(\boldsymbol{a}|\boldsymbol{v_a})P(\boldsymbol{a'}|\boldsymbol{v_a}, \boldsymbol{a}) \\ &= \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{n}\sum_{i=1}^{n}\log P(v_{a_i}) + \log P(a_i|v_{a_i}) \\ &\quad + \log P(a'_i|v_{a_i}, a_i)), \end{aligned} \tag{23}$$

where $a_1, \ldots, a_n$ are n ground-truth AU intensity values, $a'_1, \ldots, a'_n$ are corresponding measurements, and we classify $a_1, \ldots, a_n$ into two classes according to their mean value to obtain $v_{a_1}, \ldots, v_{a_n}$.

Eq. (22) is as the same as Eq. (8), therefore, Algorithm 1 is used to learn Θ_{LRBN}. Maximum log-likelihood is used to obtain parameter Θ_{BN}.

B. Network Inference

During inference phase, the learned hybrid Bayesian network model combines intensity measurements with the captured action unit dependencies to perform multiple action unit intensity estimation using probabilistic inference, according to Eq. (24):

$$\begin{aligned} &a_1, \ldots, a_n \\ &= \arg\max_{a_1,\ldots,a_n}\prod_{i=1}^{n}P(a_t|a'_i, a_i) \\ &= \arg\max_{a_1,\ldots,a_n}\prod_{i=1}^{n}P(a_i|v_{a_i})P(a'_i|a_i, v_{a_i}) \\ &\quad \times\frac{\exp(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{h}))}{\prod_j(1+\exp(w_i^T+b))\prod_j(1+\exp(d_j))}. \end{aligned} \tag{24}$$

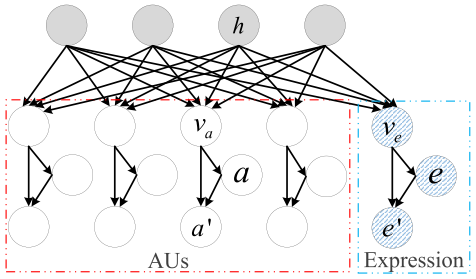


Fig. 6. The proposed AU intensity estimation enhanced by expression.

C. AU Intensity Estimation Enhanced by Expressions

Similar as AU recognition, expression dependent AU relations are expected to help AU intensity estimation. Therefore, we extend the proposed AU intensity estimation method, which is enhanced by AU dependencies, to expression-assisted AU intensity estimation method, which is enhanced by dependencies among AUs in addition to the dependencies among expressions and action units as shown in Fig. 6.

Compared with Fig. 5, in which the second layer consists solely of v_a nodes, the expression-assisted AU recognition model includes both v_a nodes and v_e nodes in the second layer. Accordingly, the lower layer contains AU intensity measurement a' as well as the expression measurement e'. If we have the ground-truth of expression intensities, the bottom layer of the expression part is three-node BNs, as the same as that of AU part. If we only have the ground-truth of expression categories, without intensities, the bottom layer of the expression part is two-node BNs, as the same as that in Fig. 4. Through incorporating expression nodes, the top two-layer LRBN uses the directed connections from hidden nodes to expression and AU nodes to thoroughly model both global relationships among multiple AUs and the relationships among AUs and expressions. The lower two-layer network consolidates facial images with the AU dependencies and the AU-expression relations, employing measurements from the images to obtain AU recognition.

The joint probability of all variables for Fig. 6 is shown in Eq. (25):

$$\begin{aligned} &P(\boldsymbol{v_a}, \boldsymbol{v_e}, \boldsymbol{a}, \boldsymbol{a'}, \boldsymbol{e}, \boldsymbol{e'}, \boldsymbol{h}) \\ &= P(\boldsymbol{a}, \boldsymbol{e}|\boldsymbol{v_a}, \boldsymbol{v_e})P(\boldsymbol{a'}, \boldsymbol{e'}|\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{v_a}, \boldsymbol{v_e}) \\ &\quad \times\frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{v_e}, \boldsymbol{h})}{\prod_j(1+\exp(d_j))}. \end{aligned} \tag{25}$$

The parameters of LRBN part and tri-BN part can be learned through Eq. (26) and Eq. (27), respectively.

$$\Theta^*_{LRBN} = \arg\max_{\Theta_{LRBN}} \frac{1}{M}\sum_{m=1}^{M} \log\Big(\sum_{\boldsymbol{h}} \frac{\exp -\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{v_e}, \boldsymbol{h})}{\prod_j (1+\exp(d_j))}\Big). \tag{26}$$

$$\Theta^*_{BN} = \arg\max_{\Theta_{BN}} \frac{1}{M}\sum_{m=1}^{M} \log P(\boldsymbol{v_a}) P(\boldsymbol{a}|\boldsymbol{v_a}) P(\boldsymbol{a'}|\boldsymbol{v_a}, \boldsymbol{a}) + \frac{1}{M}\sum_{m=1}^{M} \log P(\boldsymbol{v_e}) P(\boldsymbol{e}|\boldsymbol{v_e}) P(\boldsymbol{e'}|\boldsymbol{v_e}, \boldsymbol{e}) \tag{27}$$

Eq. (26) is as the same as Eq. (18), therefore, Eq. (26) is used to learn Θ_{LRBN}. Maximum log-likelihood is used to obtain parameter Θ_{BN}.

After parameters learning, we can inference AU intensity through Eq. (28),

$$\begin{aligned} &a_1, \ldots, a_n \\ &= \arg\max_{a_1,\ldots,a_n} \prod_{i=1}^{n} P(a_i|v_{a_i}) P(a'_i|a_i, v_{a_i}) \\ &\quad \times \max_{\boldsymbol{e}} \{P(\boldsymbol{e}|\boldsymbol{v_e}) P(\boldsymbol{e'}|\boldsymbol{e}, \boldsymbol{v_e}) \\ &\quad \times \frac{\exp(-\Gamma_{\Theta_{LRBN}}(\boldsymbol{v_a}, \boldsymbol{v_e}, \boldsymbol{h}))}{\prod_j (1+\exp(w_i^T + b)) \prod_j (1+\exp(d_j))}\}. \end{aligned} \tag{28}$$

VI. Experiments and Results

A. Experimental Conditions

To validate the proposed AU recognition and intensity estimation methods, we conducted AU recognition experiments on three benchmark databases: the Extended Cohn-Kanade (CK+) database [30], the BP4D-Spontaneous database [31] and the SEMAINE database [32]. We conduct AU intensity estimation experiments on two benchmark databases, i.e., the BP4D-Spontaneous database [31] and the UNBC-McMaster Shoulder pain Expression Archive database (PAIN) [33].

The CK+ database is made up of 593 posed facial image expression sequences taken from 123 subjects. Sequences begin with the onset frame, and end with the apex frame. Among them, 327 sequences have both annotations of expressions and AUs. All 327 of these apex frames are used in our experiments. Seven expressions (i.e., anger, contempt, disgust, fear, happiness, sadness and surprise) and 13 AUs (i.e., AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU17, AU23, AU24, AU25 and AU27) are considered. Only selected samples with frequencies larger than 10% are used.

The SEMAINE database records naturally induced facial expressions of subjects throughout a conversation. Thus far, FACS experts have coded 180 frames from eight sessions of two subjects. As in Wang *et al.* [4], 10 AUs present in at least 15 instances are used (i.e. AU1, AU2, AU4, AU5, AU6, AU7, AU12, AU17, AU25, and AU26). The SEMAINE database provides seven types of expressions: fear, anger, happiness, sadness, disgust, contempt and amusement.

The BP4D database is composed of 328 facial videos taken from 41 subjects. Each of the subjects attended eight emotion-elicitation experiments. Similar to Li *et al.* [34], Chu *et al.* [35] and Bishay and Patras [36], we use 12 AUs (i.e., AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24). For AU intensity estimation, the BP4D database only provides intensity annotations for five AUs: AU6, AU10, AU12, AU14, and AU17. We estimate intensities for all five AUs. In our experiments, all images with missing annotation and without active AUs are dropped. Finally, apex frames are adopted for AU recognition and intensity estimation respectively.

TABLE I
Distribution of AU Occurrence

	CK+	SEMAINE	BP4D
AU1	130	51	852
AU2	99	52	676
AU4	122	34	998
AU5	92	16	-
AU6	95	54	3271
AU7	79	41	3362
AU9	61	-	-
AU10	-	-	3515
AU12	80	64	2921
AU14	-	-	2772
AU15	-	-	922
AU17	115	28	1745
AU23	43	-	798
AU24	43	-	831
AU25	181	107	-
AU26	-	70	-
AU27	72	-	-

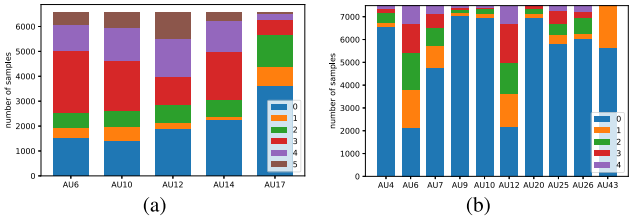


Fig. 7. Distribution of the AU intensity levels. (a) The BP4D database. (b) The PAIN database.

The PAIN database contains 200 videos from 25 subjects. A total of 48398 frames have been facial action coding system coded and active appearance model tracked. Following the same data selection criteria as Walecki *et al.* [23], the image frames with two or more active AUs were selected. For AU intensity experiments, we adopted 10 AUs (i.e., AU4, AU6, AU7, AU9, AU10, AU12, AU20, AU25, AU26, and AU43). For expression-assisted AU intensity estimation, we use the Prkachin and Solomon pain intensity (PSPI) [33] as expression-factor. Following Rudovic *et al.* [37], we discretized PSPI into six levels: 0, 1, 2, 3, 4-5, 6-15. Since the multi-labeled samples are very imbalance, we adopted multi-label balance strategy proposed by Chu *et al.* [38].

Table I shows the AU distribution on the CK+, the SEMAINE, and the BP4D databases, and Fig. 7 shows the action unit intensity distribution for the BP4D and PAIN databases.

For features, facial points provided by the database constructors are used. The feature points are normalized according

to affine transformation, such that the centers of the eyes fall on given positions for each image. For AU recognition, support vector machines (SVMs) are utilized as classifiers to collect the measurements of AU. For AU intensity estimation, we consider AU intensity estimation to be a multi-class problem, and then adopt support vector machines (SVMs) to obtain the measurements of AU intensities.

Like related works, we adopt 3-fold subject-independent cross-validation on the BP4D database. For the CK+ and SEMAINE databases, we use leave-one-subject-out cross-validation. On the PAIN database, we apply a 5-fold subject-independent cross-validation procedure. For AU recognition, F_1 score is used as the evaluation metric. AU intensity estimation uses Pearson correlation coefficient (PCC), intra-class correlation coefficient (ICC), and the mean square error (MSE) as evaluation metrics. Model selection is used to select hyper parameters.

To show the effectiveness of the proposed AU recognition methods, we conducted three kinds of experiments. First, we conduct an experiment on image-driven AU recognition. We then conduct experiments on the proposed AU recognition enhanced by AU-relations and the proposed expression-assisted AU recognition enhanced by both AU relations and expression-AU relations. The SVM is employed as the classifier recognizing the AUs from the feature points for image-driven AU recognition. As all three databases have not larger than eight expression categories, the expression labels are encoded with three binary nodes to match the action unit forms.

As in AU recognition, we conduct three kinds of experiments to validate the proposed AU intensity estimation method: the image-driven AU intensity estimation, the proposed AU intensity estimation enhanced by AU-relations, and the proposed expression-assisted AU intensity estimation enhanced by both AU relations and expression-AU relations. For the expression-assisted AU intensity estimation method, each expression corresponds to one nodes, indicating its presence/absence or larger intensity/small intensity.

As well as the above within-database experiments, we conduct cross-database AU recognition experiments using the image-driven method, the proposed method of AU recognition enhanced by AU relations, and the proposed expression-assisted AU recognition method. The common AUs and expressions between the training and the testing databases are used. Since the sample size of the BP4D database is far larger than that of the CK+ database and the SEMAINE database, we do not conduct cross-database experiments training on the small database and testing on the large database. For AU intensity estimation, we have conducted cross-database experiments using the image-driven method and the proposed AU recognition enhanced by AU relations. Similarly, as cross-database AU recognition experiments, the common AUs between the training database and the testing database are used. We have not conducted cross-database AU intensity estimation experiments using the proposed expression-assisted AU analyses, since the expression categories for the two databases are totally different. Specifically, the pain database provides pain intensity, and the BP4D database consists of eight kinds of facial expression images, i.e., happiness, sadness, surprise, embarrassment, fear, pain, anger and disgust.

B. Within-Database Experimental Results and Analyses

1) Experimental Results and Analysis on AU Recognition: The results of our AU recognition experiments on three databases can be found in Table II. From the table, we make these observations:

On all databases, the proposed AU recognition method using AU-relation modeling achieves significant improvement over the image-driven method that uses SVM. To be specific, the proposed method surpasses the image-driven method by about 3%, 5%, and 6% of average F_1 score on the CK+, SEMAINE, BP4D databases, respectively. The image-driven method does not consider AU relations, while the proposed method faithfully captures AU dependencies through LRBN. The better performance of the proposed method leveraging AU-relation modeling demonstrates that the inherent AU dependencies can be leveraged for improved AU recognition. Furthermore, the improvements on the BP4D and SEMAINE databases are larger than that on the CK+ database. The BP4D and the SEMAINE databases are made up of spontaneous facial expressions, while the CK+ database mainly consists of posed expressions. There are two possible reasons: one is that captured inherent AU dependencies may provide more benefits for spontaneous AU recognition than posed AU recognition, and the other is that the CK+ database is an easier database performance on which does not have as much room for improvement.

Secondly, the proposed expression-assisted AU recognition method outperforms both the image-driven method and the proposed AU recognition through AU-relation model, with a higher overall average F_1 score as well as higher F_1 scores for most of the AUs. Specifically, the proposed expression-assisted recognition model achieves superior performance over the proposed method of AU recognition through AU-relations, achieving a better average F_1 score by about 1% on the CK+ database and 2% on the BP4D database. These results suggest that the expression labels, which are only available during training, can capture more thoroughly AU dependencies, and thus are beneficial for AU recognition. We also find that the proposed expression-assisted AU recognition method achieves marked improvements on AU1, AU2, AU4, AU5, AU6, and AU23 on the CK+ database. These AUs are similar to the AU combinations commonly seen for surprise and fear. The improvements of AU6 and AU12 are more marked than the other AUs on the SEMAINE database. AU6 and AU12 often appear when happiness is expressed. Similarly, the improvements on the AU4, AU15 and AU23 are marked on the BP4D database. These AUs correspond to the AU combinations for anger. This provides further confirmation that the proposed expression-assisted AU recognition method successfully captures the dependencies between AUs and expressions, and effectively leverages such dependencies for AU recognition.

We also conduct hypothesis testing to further validate the superiority of the proposed methods. Specifically, 5×2 cross-validation paired t-test [39] are conducted to verify the improvement of the proposed model enhanced by

TABLE II
F_1 SCORE OF AU OCCURRENCE RECOGNITION

Method	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14	AU15	AU17	AU23	AU24	AU25	AU26	AU27	Avg.
CK+																		
Image-driven	0.84	0.87	0.75	0.80	0.76	0.63	0.94		0.86			0.88	0.70	0.45	0.96		0.89	0.79
Model-based	0.88	0.90	0.78	0.81	0.76	**0.65**	**0.94**	-	0.87	-		**0.88**	0.72	**0.56**	0.97	-	**0.91**	0.82
Expression	**0.93**	**0.93**	**0.82**	**0.84**	**0.84**	0.63	0.93		**0.89**			0.86	**0.75**	0.53	**0.97**		0.90	**0.83**
SEMAINE																		
Image-driven	0.73	0.74	0.56	0.37	0.64	0.56			0.66			0.33			0.88	0.63		0.61
Model-based	0.83	**0.83**	0.58	0.40	0.67	0.61	-		0.76	-		0.37	-		**0.90**	**0.67**	-	0.66
Expression	**0.84**	0.82	**0.58**	**0.40**	0.69	**0.61**			**0.77**			**0.37**			0.90	0.67		**0.66**
BP4D																		
Image-driven	0.29	0.21	0.50		0.84	0.85		0.88	0.92	0.69	0.36	0.65	0.37	0.59				0.60
Model-based	**0.36**	0.32	0.53	-	0.90	0.88	-	0.89	0.93	0.76	0.44	0.70	0.37	0.63		-		0.64
Expression	0.34	**0.32**	**0.62**		**0.91**	**0.89**		**0.90**	**0.94**	**0.76**	**0.46**	**0.70**	**0.39**	**0.64**				**0.66**

TABLE III
p-VALUES OVER AVERAGE F_1 SCORE FOR AU RECOGNITION

	SVM vs. Model	SVM vs. Exp	Model vs. Exp
CK+	**1.46e-02**	**6.09e-05**	1.96e-01
SEMAINE	**2.31e-16**	**4.95e-17**	1.61e-01
BP4D	**2.24e-10**	**8.28e-13**	**1.70e-03**

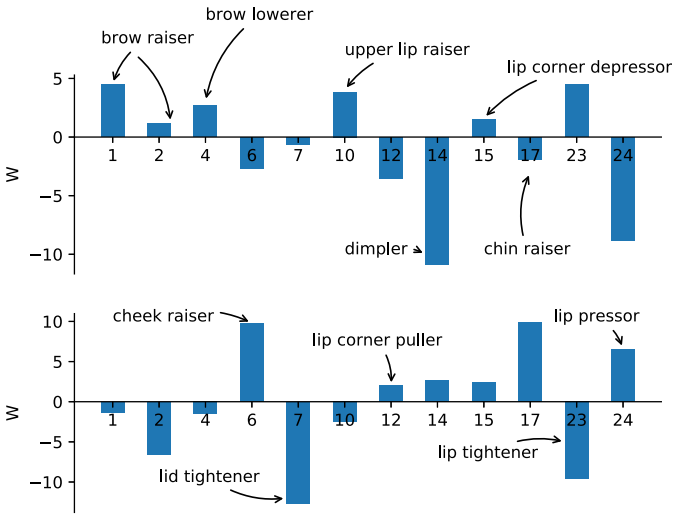


Fig. 8. Semantic relationship captured by two latent nodes of LRBN for AU recognition.

AU relations to images-based method (i.e., SVM vs. Model) and the improvement of the proposed model enhanced by AU relations and expression over the image-based method or the proposed model enhanced by AU relations (i.e., SVM vs. Exp and Model vs. Exp). As shown in Table III, most of the p-values for F_1 score are less than 0.05. It suggests that the improvement is significant for the proposed AU recognition methods.

To show the effectiveness of the proposed LRBN in capturing AU relations, we graphically illustrated the captured AU dependencies in Fig. 8. As we discuss in Section III, each of the latent nodes is able to capture a specific pattern. This pattern is measured by weights W_{ij} between the latent nodes and AUs. A larger weight indicates a high probability of occurrence. A smaller weight is indicative of a higher probability of absence. The figure shows the first latent node, which encodes a pattern for a person who is likely to simultaneously "raise brow," "lower brow," and "raise lip," but is probably not going to "dimple." This likely represents AU relations for a negative emotion like sadness, fear, or anger. The second latent node encodes a pattern for a person who is likely to "raise cheek," "pull lip corner," and "press lip," but is less likely to either "tighten lid" or "tighten lip." This could represent AU relations for a positive emotion, i.e., happiness. These learned relations show that the proposed model is able to effectively capture global AU relationships. Thus, proposed models obtain better performance of AU recognition.

To analyze the limitations of the proposed method, we manually check the misclassified samples. Take the BP4D database as an example, AU7, AU10 and AU12 occur simultaneously for a sample. The sample is predicted correctly by image-based method, but two proposed methods predict that AU6, AU7, AU10 and AU12 occur simultaneously. We find that the ground-truth AU combination is a small probability event. Specifically, the probability $P(AU6 = 0|AU7 = 1, AU10 = 1, AU12 = 1)$ is less than 0.05. And given other AUs, the probability of AU6 = 1 is around 350 times that of AU6 = 0. Therefore, the AU combination is predicted to a more likely combination by the proposed methods. It suggests that proposed methods may ignore the small probability event during exploiting the AU relations. In addition, the proposed methods may not perform well on datasets that have very different AU and expression relationships and biases from the training dataset.

2) Experimental Results and Analysis on AU Intensity Estimation: The results of the experiments on AU intensity estimation are shown in Table IV. The table shows the following:

First, compared to the image-driven method employing SVM, the proposed AU intensity estimation model leveraging AU relations modeling achieves superior performance on the

TABLE IV
AU INTENSITY ESTIMATION RESULTS

	Method	AU4	AU6	AU7	AU9	AU10	AU12	AU14	AU17	AU20	AU25	AU26	AU43	Avg.
							BP4D							
PCC	Image-driven	-	0.70	-	-	0.68	0.83	0.23	0.51	-	-	-	-	0.59
	Model-based		0.73			0.71	0.85	0.43	0.58					0.66
	Expression		**0.76**			**0.71**	**0.86**	**0.53**	**0.71**					**0.71**
ICC	Image-driven	-	0.70	-	-	0.68	0.83	0.23	0.51	-	-	-	-	0.59
	Model-based		0.73			0.71	0.85	0.43	0.57					0.66
	Expression		**0.76**			**0.71**	**0.85**	**0.52**	**0.69**					**0.70**
MSE	Image-driven	-	1.23	-	-	1.33	1.10	2.11	1.22	-	-	-	-	1.40
	Mode-based		1.20			1.25	1.02	1.79	1.11					1.27
	Expression		**1.10**			**1.14**	**1.02**	**1.61**	**0.94**					**1.16**
							PAIN							
PCC	Image-driven	0.60	0.64	0.60	0.29	0.70	0.70	-	-	0.36	0.60	0.64	0.69	0.58
	Model-driven	0.69	0.84	0.47	0.66	0.79	0.74			**0.64**	0.59	**0.89**	0.71	0.70
	Expression	**0.77**	**0.88**	**0.63**	**0.69**	**0.82**	**0.86**			0.56	**0.69**	0.69	**0.77**	**0.74**
ICC	Image-driven	0.58	0.63	0.58	0.29	0.70	0.69	-	-	0.22	0.60	0.64	0.65	0.61
	Model-driven	0.63	0.82	0.46	0.52	0.75	0.71			**0.62**	0.54	**0.88**	0.67	0.66
	Expression	**0.75**	**0.87**	**0.63**	**0.61**	**0.81**	**0.86**			0.49	**0.65**	0.68	**0.76**	**0.71**
MSE	Image-driven	0.48	1.34	1.41	0.53	0.21	1.29	-	-	0.30	0.96	0.78	**0.08**	0.74
	Model-driven	0.37	0.64	1.15	0.24	0.15	1.19			0.27	0.87	**0.25**	0.11	0.52
	Expression	**0.28**	**0.46**	**1.09**	**0.21**	**0.12**	**0.60**			**0.21**	**0.70**	0.62	0.11	**0.44**

TABLE V
p-VALUES OVER AVERAGE PCC, ICC AND MSE FOR AU INTENSITY ESTIMATION

		SVM vs. Model	SVM vs. Exp	Model vs. Exp
BP4D	PCC	**1.36E-02**	**9.97E-08**	**2.21E-02**
	ICC	**2.95E-04**	**1.59E-07**	**5.13E-04**
	MSE	**1.61E-02**	**1.42E-02**	8.77E-02
PAIN	PCC	**1.40E-03**	**2.05E-05**	**1.91E-03**
	ICC	**8.01E-05**	**1.47E-05**	**2.82E-02**
	MSE	**5.32E-03**	**1.80E-03**	**3.00E-02**

BP4D database and the PAIN database. On the BP4D database, the proposed model has an improved performance over the image-driven method by about 7% over average PCC and 7% over average ICC. In addition, the proposed model shows a 0.13 decrease in average MSE compared to the image-driven method. On the PAIN database, the proposed model achieves an improvement of 12% over average PCC and 5% over average ICC in comparison to the image-driven model. Besides, the proposed model shows a decrease of about 0.22 in average MSE when compared to the image-driven method. Unlike image-driven method, which ignores AU dependencies, the proposed AU intensity estimation through AU relation modeling completely captures AU dependencies through LRBN. The superiority of the proposed method over the image-driven method suggests that the inherent AU dependencies are crucial for AU intensity estimation.

Secondly, the proposed AU intensity estimation enhanced by expressions outperforms the image-driven method as well as the proposed AU intensity estimation model leveraging AU relations, with higher PCC, ICC and lower MSE for most AUs. On the BP4D database, the expression-assisted model achieves superior performance over the AU-relation model by about 5% over average PCC and 7% over average ICC, separately. In addition, the expression-assisted model decreases the AU-relation model by about 0.11 in average MSE. On the PAIN database, the expression-assisted model outperforms the AU-relation model by about 4% over average PCC and 5% over average ICC, respectively. Besides, the proposed model shows a decrease of about 0.08 for average MSE when compared to the AU-relation model. All these improvements demonstrate that AU intensity estimation can be further facilitated through the semantic relationship between AUs and expressions.

Thirdly, we can observe that the proposed two AU intensity estimation models have marked improvements for AU14 and AU17 on the BP4D database. For image-driven method, the PCC and ICC of the two AUs are lower than those of other three AUs. As shown in Fig. 7(a), the distribution of AU14 and AU17 is more imbalanced than others. Such imbalanced data distribution may cause their worse performance in image-driven method. While the proposed methods can successfully capture AU dependencies, and leverage the captured AU dependencies to improve their performances. On the PAIN database, the proposed two AU intensity estimation models have marked improvements for AU9, i.e. nose wrinkler, compared to image-driven method. Since we use feature points as the features, it is no wonder that the image-driven method cannot recognize AU9 well. AU9 is one of the primary AUs in expressing pain expressions as indicated in the PSPI score [33], PSPI = AU4 + max(AU6,AU7) + max(AU9,AU10) + AU43. By capturing global AU dependencies and AU-expression dependencies, the proposed AU intensity estimation models can obtain marked improvement for AU9 recognition.

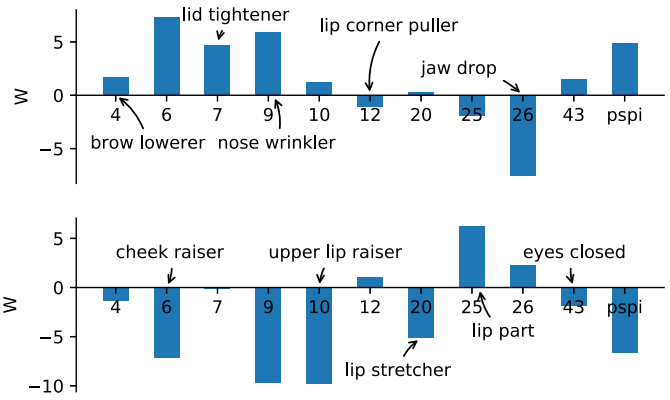


Fig. 9. Semantic relationship captured by LRBN for AU intensity estimation on the PAIN database.

For AU intensity estimation experiments, we also conduct the 5×2 cross validation paired t-test. As shown in Table V, all of the p-values are less than 0.05 for PCC and ICC, and most of the p-values are less than 0.05 for MSE. It illustrates that the improvement is significant for the proposed AU intensity estimation methods.

To show the effectiveness of the proposed LRBN in capturing AU relations for AU intensity estimation, we graphically illustrated the captured AU dependencies in Fig. 9. The first latent node encodes a pattern for a person who is likely to simultaneously exercise the AUs for "brow lowerer," "cheek raise," "lid tightener," and "nose wrinkle," which may represent AU relations for sharp pain. The second latent node encodes the pattern for "lip corner puller" and "lip part," which may represent AU relations for painless. These learned relations show that the proposed model effectively captures the global AU dependencies among AU and expression.

3) Comparison With Related Works: To further demonstrate the superiority of the proposed AU recognition method, we compare our proposed method to state-of-the-art works leveraging AU relationships for AU recognition.

We compare our proposed method to MC-LVM [12], BN [17], and HRBM+ [4] on the CK+ database. Since BN [12] and HRBM+ [17] do not offer results from experiments on the SEMAINE database, we instead choose to compare our proposed method to HRBM+ [4]. Most AU recognition works on the BP4D database employed deep networks. Therefore, we opt to compare our proposed method to four recent deep works: Wu *et al.* [5], Li *et al.* [34], Chu *et al.* [35] and Bishay and Patras [36]. Since the experimental conditions of the deep works and ours are not exactly the same, the comparisons are only for reference. The comparisons to the related works on AU recognition are illustrated in Table VI.

Furthermore, we validate the superiority of the proposed AU intensity estimation method over recent state-of-the-art methods leveraging AU relationships for AU intensity estimation. For the BP4D database, we compare the proposed method with Valstar *et al.* [41], Gudi *et al.* [42], Wang *et al.* [17] and Linh Tran *et al.* [43]. Valstar *et al.* [41] provided a baseline intensity estimation on the BP4D database. As mentioned in Section II, Wang *et al.* [17] used BN to model AU dependencies and AU-expression dependencies. Gudi *et al.* [42], and Linh Tran *et al.* [43] are recent deep AU intensity estimation works. For the PAIN database, our proposed method is compared to MLT+BN [19], LT-all [22], COR-LIT [23] and cs-CORF [24]. All of the compared work leveraged AU dependencies for intensity estimation. The comparisons with related works of AU intensity estimation are shown in Table VII.

TABLE VI
COMPARISON WITH RELATED WORKS OVER AU OCCURRENCE RECOGNITION

Database	Method	F_1 score
CK+	MC-LVM [12]	0.781
	BN [17]	0.800
	HRBM+ [4]	0.8244
	Ours	**0.83**
SEMAINE	HRBM+ [4]	0.6079
	Jiang *et al.* [40]	0.6083
	Ours	**0.66**
BP4D	Wu *et al.* [5]	0.491
	Li *et al.* [34]	0.559
	Chu *et al.* [35]	0.532
	Bishay *et al.* [36]	0.627
	Ours	**0.66**

TABLE VII
COMPARISON WITH RELATED WORKS OVER AU INTENSITY ESTIMATION

Database	Method	PCC	ICC	MSE
BP4D	Valstar *et al.* [41]	0.554	0.592	1.401
	CNN [42]	0.621	0.613	1.181
	BN [17]	0.632	0.600	-
	2DC [43]		0.66	**0.95**
	Ours	**0.71**	**0.70**	1.16
PAIN	MLT+BN [19]	0.28	0.38	**0.36**
	LT-all [22]	-	0.23	0.60
	COR-LIT [23]	0.33	0.37	1.10
	cs-CORF [24]	0.617	-	-
	Ours	**0.71**	**0.74**	0.44

Table VI shows that the proposed AU recognition model enhanced by expressions achieves the best performance on every tested database. The average F_1 score for the proposed method is about 4.9% higher than MC-LVM [12], 3.0% higher than BN [17] and 0.6% higher than HRBM+ [4] for the CK+ database. As mentioned in Section II-A, MC-LVM [12] used constrains to represent co-occurrence AU dependencies, BN [17] adopted BN to capture pair wised AU dependencies, and HRBM+ [4] employed RBM to model global AU dependencies and AU-expression dependencies. Although all the compared works take advantage of AU dependencies for AU recognition, their captured AU dependencies are less thoroughly and faithfully than the proposed method. Therefore, the proposed method achieves better performance. On the SEMAINE database, the average F_1 score of proposed method is about 5.3% higher than [4] and 5.2% higher than [40]. The superior performance of the proposed method to HRBM+ further demonstrates the more powerful data representation ability of LRBN than other generative models,

TABLE VIII
CROSS-DATABASE EXPERIMENTAL RESULTS OF AU RECOGNITION

Method	AU1	AU2	AU4	AU5	AU6	AU7	AU12	AU17	AU23	AU24	AU25	Avg.
BP4D → CK+												
Image-driven	0.68	0.62	0.64		0.62	**0.43**	0.64	0.81	0.47	0.46		0.60
Model-based	**0.73**	**0.63**	0.64	-	0.63	0.42	0.65	**0.83**	**0.51**	0.46	-	0.61
Expression	0.68	0.62	0.70		**0.64**	0.42	**0.75**	**0.83**	0.51	**0.47**		0.63
BP4D → SEMAINE												
Image-driven	0.67	0.70	0.53		**0.57**	**0.46**	0.61	0.40				0.56
Model-based	0.67	**0.73**	**0.60**	-	0.56	0.44	**0.62**	0.40		-		**0.57**
Expression	**0.70**	**0.73**	0.56		0.56	0.43	**0.62**	**0.42**				**0.57**
CK+ → SEMAINE												
Image-driven	0.54	0.55	0.35	0.24	0.44	0.20	0.38	**0.56**			**0.91**	0.46
Model-based	**0.58**	**0.59**	**0.40**	0.24	**0.48**	0.20	**0.45**	0.54		-	0.89	**0.49**
Expression	0.54	0.58	**0.40**	0.24	**0.48**	0.20	0.44	**0.56**			**0.91**	0.48
SEMAINE → CK+												
Image-driven	0.66	0.80	0.51	0.10	0.68	0.27	**0.82**	0.47			0.94	0.58
Model-based	0.69	0.80	0.51	0.10	0.70	0.30	0.76	0.50		-	0.94	**0.59**
Expression	**0.74**	**0.82**	**0.53**	0.10	0.69	**0.32**	0.71	0.48			0.94	**0.59**

such as RBM. The average F_1 score of the proposed method is about 16.9% higher than [5], 10.1% higher than [34], 12.8% higher than [35] and 3.3% higher than [36] for the BP4D database. Although these deep AU recognition works can learn better facial representations, they do not capture AU dependencies from the label-level. These results strongly suggest the superiority of proposed method. This suggests that the proposed method is superior to other state-of-the-art works.

Table VII demonstrates that our proposed expression-assisted AU intensity estimation method outperforms all related AU intensity estimation works enhanced by AU dependencies. To be specific, on the BP4D database, the average PCC and ICC of proposed method are about 7.8% and 10.0% higher than Wang *et al.* [17]. On the PAIN database, the proposed method outperforms all the compared works with higher PCC and ICC as well as lower MSE, except for MSE of MLT+BN [19]. Wang *et al.* [19] did not use AU43, which does not have intensity annotation, but occurrence state. While the proposed method recognized AU43 in addition to estimate intensities of other 10 AUs as shown in Table IV. The recognition of AU43 increases the average MSE. As mentioned in Section II, Wang *et al.* [19] adopted the constraints among multiple tasks as representations of AU relations. Kaltwang *et al.* [22], Walecki *et al.* [23] and Rudovic *et al.* [24] adopted generative latent tree, conditional random field, and conditional ordinal random field to capture AU dependencies respectively. Wang *et al.* [17] adopted BN to model dependencies among AU intensities and expressions. The better performance of the proposed method in AU intensity estimations further demonstrates the superiority of the LRBN in capturing AU dependencies over state-of-the-art work.

Compared with current deep AU intensity estimation works, the proposed work achieves better performance with higher ICC for the BP4D database. Specifically, on the proposed method, the average ICC is about 8.7% higher than CNN [42], and 4.0% higher than 2DC [43]. Although the deep AU intensity estimation works take advantage of the strength of deep network in representation learning, they do not explicitly explore AU dependencies from the label-level. Our superior performance further demonstrates the importance of AU dependencies for intensity estimation.

C. Cross-Database Experimental Results and Analyses

1) Experimental Results and Analysis on AU Recognition: The cross-database experimental results of AU recognition are shown in Table VIII.

From Table VIII, we can find that the proposed AU-relation recognition method and the proposed expression-assisted AU recognition method outperform the image-based method with higher F_1 scores in most cases. It demonstrates that the proposed methods are able to effectively leverage the captured AU relations for AU recognition. Compared Table VIII with Table II, we can find that the improvements of the proposed methods to the image-based method for the cross-database experiments are smaller than those for the within database experiments. Specifically, for within database experiments, the proposed methods enhanced by AU relations outperforms the image-based methods by about 2%-6% on the three databases. But for the cross-database experiments, the improvements are less than 3%. This may be due to the database bias.

Furthermore, from Table VIII, we find that when comparing our proposed AU-relation enhanced model for AU recognition, the proposed AU recognition method enhanced by expression does not improve the performance of AU recognition effectively. The reason may be that the expression induced method for the three databases are totally different. Specifically, the seven basic emotion categories are posed and they are labeled according to AU combination on the CK+ database. For the SEMAINE database, the seven expressions come from interactions between users and agent. For the BP4D database,

TABLE IX
CROSS-DATABASE EXPERIMENTAL RESULTS OF AU INTENSITY ESTIMATION

	Method	AU6	AU10	AU12	Avg.
			BP4D → PAIN		
ICC	Image-driven	0.36	0.06	0.38	0.26
	Model-based	**0.36**	**0.08**	**0.49**	**0.31**
PCC	Image-driven	**0.38**	0.21	0.40	0.33
	Model-based	0.37	**0.26**	**0.50**	**0.38**
MSE	Image-driven	2.87	6.37	2.81	4.02
	Model-based	**2.40**	**5.06**	**2.30**	**3.25**
			PAIN → BP4D		
ICC	Image-driven	0.39	0.02	0.15	0.19
	Model-based	**0.44**	**0.05**	**0.22**	**0.23**
PCC	Image-driven	0.51	0.10	0.19	0.27
	Model-based	**0.55**	**0.15**	**0.23**	**0.31**
MSE	Image-driven	**2.20**	7.75	3.22	4.39
	Model-based	2.23	**7.28**	**2.27**	**3.93**

the spontaneous emotions are elicited through eight designed emotion-induced tasks.

2) Experimental Results and Analysis on AU Intensity Estimation: From Table IX, we find that the proposed AU intensity estimation enhance by AU relations performs better than the image-based method, with higher ICC, higher PCC and lower MSE in most cases. The better performance demonstrates that the proposed model successfully captures the anatomical relationships among AUs for AU analyses. Due to the database biases, the improvement on cross-database experiments is smaller than that on within-database experiments.

VII. CONCLUSIONS

This paper proposes a novel approach for multiple AU recognition, and a novel intensity estimation method for multiple AUs that captures the global dependencies among action units. Hierarchical models are employed with a hybrid Bayesian network. To be specific, the upper two layers are an LRBN model capturing global dependencies among the action units. The lower layers are Bayesian networks. They connect ground-truth AU labels with their respective measurements. The hierarchical Bayesian model leverages the dependencies between expressions and AUs. It is enhanced during training with facial expression labels to further improve AU intensity estimation and recognition performance. The results of our experiments on three benchmark databases (for AU recognition) and two benchmark databases (for AU intensity estimation) show that the proposed methods are able to effectively capture relationships among AUs and between AUs and expressions, thereby improving AU intensity estimation and multiple AU recognition.

REFERENCES

[1] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA, USA: Consulting Psychologists, 1978.

[2] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.

[3] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2559–2573, Jul. 2013.

[4] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3304–3311.

[5] S. Wu, S. Wang, B. Pan, and Q. Ji, "Deep facial action unit recognition from partially labeled data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3971–3979.

[6] S. Wang, S. Wu, G. Peng, and Q. Ji, "Capturing feature and label relations simultaneously for multiple facial action unit recognition," *IEEE Trans. Affect. Comput.*, to be published.

[7] S. Nie, Y. Zhao, and Q. Ji, "Latent regression Bayesian network for data representation," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3494–3499.

[8] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2017.2731763.

[9] Y. Zhu, S. Wang, L. Yue, and Q. Ji, "Multiple-facial action unit recognition by shared feature learning and semantic relation modeling," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1663–1668.

[10] X. Zhang and M. H. Mahoor, "Simultaneous detection of multiple facial action units via hierarchical task structure learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1863–1868.

[11] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2207–2216.

[12] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3792–3800.

[13] A. Ruiz, J. Van de Weijer, and X. Binefa, "From emotions to action units with hidden and semi-hidden-task learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3703–3711.

[14] P. Gosselin, G. Kirouac, and F. Y. Doré, "Components and recognition of facial expression in the communication of emotion by actors," *J. Pers. Soc. Psychol.*, vol. 68, no. 1, pp. 83–96, Jan. 1995.

[15] K. R. Scherer and H. Ellgring, "Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal?" *Emotion*, vol. 7, no. 1, pp. 113–130, Mar. 2007.

[16] J. Wang, S. Wang, and Q. Ji, "Facial action unit classification with hidden knowledge under incomplete annotation" in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 75–82.

[17] S. Wang, Q. Gan, and Q. Ji, "Expression-assisted facial action unit recognition under incomplete Au annotation," *Pattern Recognit.*, vol. 61, pp. 78–91, Jan. 2017.

[18] J. Nicolle, K. Bailly, and M. Chetouani, "Facial action unit intensity prediction via hard multi-task metric learning for kernel regression," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 6, May 2015, pp. 1–6.

[19] S. Wang, J. Yang, Z. Gao, and Q. Ji, "Feature and label relation modeling for multiple-facial action unit classification and intensity estimation," *Pattern Recognit.*, vol. 65, pp. 71–81, May 2017.

[20] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji, "A unified probabilistic framework for measuring the intensity of spontaneous facial action units," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[21] G. Sandbach, S. Zafeiriou, and M. Pantic, "Markov random field structures for facial action unit intensity estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 738–745.

[22] S. Kaltwang, S. Todorovic, and M. Pantic, "Latent trees for estimating intensity of facial action units," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 296–304.

[23] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Copula ordinal regression for joint estimation of facial action unit intensity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4902–4910.

[24] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 944–958, May 2015.

[25] L. Hao, S. Wang, G. Peng, and Q. Ji, "Facial action unit recognition augmented by their dependencies," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 187–194.

[26] M. P. Wellman and M. Henrion, "Explaining 'explaining away,'" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 3, pp. 287–292, Mar. 1993.
[27] H. Robbins and S. Monro, "A stochastic approximation method," vol. 22, no. 3, pp. 400–407, Sep. 1951.
[28] A. Yuille, "The convergence of contrastive divergences," Ph.D. dissertation, Dept. Statist., Univ. California Los Angeles, Los Angeles, CA, USA, 2006.
[29] Z. Gao, S. Wang, and Q. Ji, "Multiple aesthetic attribute assessment by exploiting relations among aesthetic attributes," in *Proc. 5th Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 575–578.
[30] P. Lucey, J. F. 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.
[31] X. Zhang *et al.*, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.
[32] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan./Mar. 2012.
[33] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. Face Gesture*, Mar. 2011, pp. 57–64.
[34] W. Li, F. Abtahi, Z. Zhu, and L. Yin. (May 2017). "EAC-net: A region-based deep enhancing and cropping approach for facial action unit detection," [Online]. Available: https://arxiv.org/abs/1702.02925
[35] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, vol. 4, Jun. 2017, pp. 25–32.
[36] M. Bishay and I. Patras, "Fusing multilabel deep networks for facial action unit detection," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Jun. 2017, pp. 681–688.
[37] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive conditional ordinal random fields for facial action intensity estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 492–499.
[38] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning facial action units with spatiotemporal cues and multi-label sampling," *Image Vis. Comput.*, 2018.
[39] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[40] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. Face Gesture*, Mar. 2011, pp. 314–321.
[41] M. F. Valstar *et al.*, "FERA 2015—Second Facial Expression Recognition and Analysis challenge," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 6, May 2015, pp. 1–8.
[42] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, "Deep learning based FACS action unit occurrence and intensity estimation," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 6, May 2015, pp. 1–5.
[43] D. L. Tran *et al.*, "Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2017, pp. 3190–3199.

Shangfei Wang (SM'15) received the B.S. degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996, and the M.S. degree in circuits and systems and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1999 and 2002, respectively. From 2004 to 2005, she was a Post-Doctoral Research Fellow with Kyushu University, Japan. From 2011 to 2012, she was a Visiting Scholar with the Rensselaer Polytechnic Institute, Troy, NY, USA. She is currently an Associate Professor with the School of Computer Science and Technology and the School of Data Science, USTC. She has authored or co-authored over 90 publications. Her research interests cover affective computing and probabilistic graphical models. She is a member of the ACM.

Longfei Hao received the B.S. degree in computer science from Anhui University in 2016. He is currently pursuing the M.S. degree in computer science with the University of Science and Technology of China, Hefei, China. His research interest is in affective computing.

Qiang Ji (F'15) received the Ph.D. degree in electrical engineering from the University of Washington. From 2009 to 2010, he was the Program Director of the National Science Foundation (NSF), Arlington, VA, USA, where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions at the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL, USA, the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, the Department of Computer Science, University of Nevada at Reno, Reno, NV, USA, and the Air Force Research Laboratory, Rome, NY, USA. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, where he is also the Director of the Intelligent Systems Laboratory.

He has published over 230 papers in peer-reviewed journals and conferences. His research interests are in computer vision, probabilistic graphical models, and machine learning and their applications in various fields. He is a fellow the IAPR. He is a program committee member of numerous international conferences/workshops. He received multiple awards for his work. He has served as the general chair, the program chair, and the technical area chair for numerous international conferences/workshops. He is currently an editor of several his research-related IEEE and international journals.