Differentiating Between Posed and Spontaneous Expressions with Latent Regression Bayesian Network

Quan Gan,[1*] **Siqi Nie,**[2*] **Shangfei Wang,**[1†] **Qiang Ji**[2]
[1]School of Computer Science and Technology, University of Science and Technology of China
[1]{gqquan@mail., sfwang@}ustc.edu.cn
[2]Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute
[2]{nies, jiq}@rpi.edu

Abstract

Spatial patterns embedded in human faces are crucial for differentiating posed expressions from spontaneous ones, yet they have not been thoroughly exploited in the literature. To tackle this problem, we present a generative model, i.e., Latent Regression Bayesian Network (LRBN), to effectively capture the spatial patterns embedded in facial landmark points to differentiate between posed and spontaneous facial expressions. The LRBN is a directed graphical model consisting of one latent layer and one visible layer. Due to the "explaining away" effect in Bayesian networks, LRBN is able to capture both the dependencies among the latent variables given the observation and the dependencies among visible variables. We believe that such dependencies are crucial for faithful data representation. Specifically, during training, we construct two LRBNs to capture spatial patterns inherent in displacements of landmark points from spontaneous facial expressions and posed facial expressions respectively. During testing, the samples are classified into posed or spontaneous expressions according to their likelihoods on two models. Efficient learning and inference algorithms are proposed. Experimental results on two benchmark databases demonstrate the advantages of the proposed approach in modeling spatial patterns as well as its superior performance to the existing methods in differentiating between posed and spontaneous expressions.

Introduction

Distinguishing between posed and spontaneous expressions is crucial for human-computer interaction, even human-human interaction, since spontaneous expressions reveal one's real emotions, while posed expressions may disguise one's inner feelings.

The main components of current work on posed and spontaneous expression distinction consists of feature extraction and classification. For feature extraction, most research proposes temporal features and spatial features specially designed for differentiating posed expressions from spontaneous ones. Temporal features include duration, amplitude, speed, acceleration, symmetry and trajectory (Cohn and Schmidt 2004) (Valstar et al. 2006) (Dibeklioglu, Salah, and Gevers 2015) (Seckington 2011), and spatial features consist of distance and angular features (Dibeklioglu et al. 2010). In addition to defining posed vs spontaneous expression specified features, some research adopts commonly used features for expression recognition, such as Gabor wavelet features (Littlewort, Bartlett, and Lee 2009), Completed local binary patterns from Three Orthogonal Planes (CLBP-TOP) (Pfister et al. 2011a), scale-invariant feature transform (SIFT) appearance features and facial animation parameters (FAP) geometric features (Zhang, Tjondronegoro, and Chandran 2011). Instead of using hand-crafted features, Gan *et al.* (Gan et al. 2015) proposed to learn features using a deep Boltzmann machine. After feature extraction, both static and dynamic machine learning methods have been investigated to distinguish posed expressions from spontaneous expressions. The static classifiers include linear discriminant analysis (Cohn and Schmidt 2004), support vector machines (Littlewort, Bartlett, and Lee 2009), Adaboost (Littlewort, Bartlett, and Lee 2009), gentle Boost and relevance vector machines (Valstar et al. 2006). The dynamic classifiers include hidden Markov models (Dibeklioglu et al. 2010) and dynamic Bayesian networks (Seckington 2011). The static classifiers capture the mapping between extracted features and expressions ignoring the dynamic aspects of expressions, while the dynamic ones can model the temporal dynamics. All these research demonstrates the progress in distinguishing posed and spontaneous expressions. However, most current works employ different features and classifiers for posed and spontaneous expression distinction, without explicitly capturing spatial patterns embedded in posed and spontaneous expression respectively, and leverage such spatial patterns for posed and spontaneous expression distinction. We call them feature-driven methods.

Behavior research indicates that posed and spontaneous expressions are different from each other in both temporal and spatial patterns. Spatial patterns mainly consist of the movement of facial muscles that show up as the occurrence of facial action units (AUs). For example, for spontaneous disgust, the three most frequently observed AUs are AU6, AU7 and AU10, while for posed disgust, they are AU4, AU7 and AU17. For posed sadness, the three most frequently observed AUs are AU4, AU7, AU17, and this pattern doesn't apply to spontaneous sadness (Namba et al. 2016). Both zygomatic major and orbicularis oculi are contracted dur-

[*] The two authors contributed equally to this paper.

[†] This is the corresponding author.

ing spontaneous smiles, while only zygomatic major is contracted for posed smiles (Ekman and Friesen 1982); the contraction of zygomatic major is more likely to occur asymmetrically for posed smiles than spontaneous ones (Ekman, Hager, and F. 1981). These observations from nonverbal behavior research prove that spatial patterns embedded in posed and spontaneous expressions are crucial for differentiating posed expressions from spontaneous ones.

Only recently, Wang *et al.* (Wang et al. 2015) proposed multiple Bayesian networks (BN) to capture posed and spontaneous spatial facial patterns respectively given gender and expression categories. We call it a model-based method. However, due to the first-order Markov assumption of their model, only the the local dependencies among geometric features are captured. Compared with their BN model, restricted Boltzmann machines (RBM) can model higher-order dependencies among random variables by introducing a layer of latent units (Hinton 2010). Wu (Wu and Wang 2016) proposed to use restricted Boltzmann machines to explicitly model complex joint distributions over feature points, i.e., spatial patterns, embedded in posed and spontaneous expressions respectively. Specifically, they constructed multiple RBMs to model spatial patterns from facial geometric features. Their experimental results demonstrate the effectiveness of RBMs in modelling global spatial patterns as well as its superior posed and spontaneous expression distinction performance over existing approaches.

Although RBM can effectively capture global dependencies among visible units through introducing hidden units, hidden units are independent to each other given visible units. Introducing dependencies among hidden units will increase the model power in explaining the patterns embedded in the visible units. Unlike RBM, which is an undirected latent variable model, LRBN is a directed latent variable model. It can better represent the visible units through directed links among hidden units and visible units. Due to the strong ability for data representation, LRBN can be applied to wide range of domains for both data representation and classification. Therefore, in this paper, we propose employing the LRBN to effectively capture the high-order and global dependencies among facial geometric features.

During training, we train two LRBN models using posed and spontaneous expression data respectively. The visible variables of LRBN represent the displacements of feature points. Thus, spatial patterns of posed and spontaneous expressions are captured by the dependences among visible nodes and the dependencies among hidden nodes of the two LRBN models respectively. During testing, the images are assigned the expression type labels whose models have the maximum likelihood. Experimental results on two benchmark databases show the advantage of our proposed method and demonstrate the strong ability of the LRBN model to capture spatial patterns which are helpful to differentiate between posed and spontaneous expressions.

Proposed method

The LRBN is a special kind of Bayesian Network, consisting of one visible layer and one latent layer, every latent variable connects to every visible variable with a directed edge as shown in Fig. 1 .

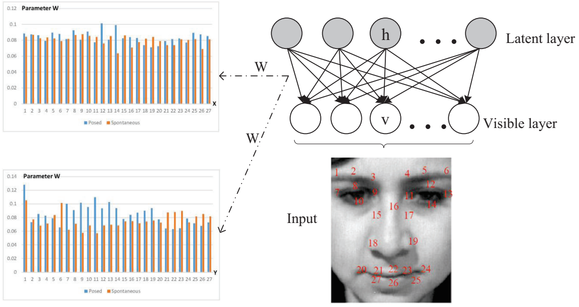


Figure 1: The framework of capturing spatial patterns.

According to the chain rule in Bayesian Networks, the joint probability of all variables is factorized into the product of prior probabilities and conditional probabilities as shown in Eq. 1,

$$P(\boldsymbol{x}, \boldsymbol{h}) = \prod_{j=1}^{n_h} P(h_j) \prod_{i=1}^{n_d} P(x_i|\boldsymbol{h}) \,. \tag{1}$$

The prior probability for a latent variable h_j is defined as Eq. 2,

$$P(h_j = 1) = \mathrm{sigm}(d_j) \,, \tag{2}$$

where $\mathrm{sigm}(d_j) = 1/(1 + \exp(-d_j))$ is the sigmoid function, and d_j is the parameter. This formulation is essentially a Bernoulli distribution.

The conditional probability of a visible variable given all the latent variables is defined as a linear Gaussian, as shown in Eq. 3,

$$P\left(x_i|\boldsymbol{h}\right) \sim \mathcal{N}\left(\boldsymbol{w}_i^T\boldsymbol{h} + b_i,\, \sigma_i\right) \,, \tag{3}$$

where the mean is a linear combination of the values of the latent variables. w_{ij} is the weight for node h_j and x_i; b_i is a constant term; and σ_i is the standard deviation. Thus, the LRBN can be viewed as a mixture of Gaussian with the number of components exponential in the number of latent variables.

Plugging in the prior distributions and conditional distributions, the joint distribution of visible variables and hidden variables has the following formulation,

$$\begin{aligned} P_\Theta(\boldsymbol{x}, \boldsymbol{h}) &= \prod_j \frac{\exp(d_j h_j)}{1 + \exp(d_j)} \prod_i \mathcal{N}(x_i : \boldsymbol{w}_i^T\boldsymbol{h} + b_i,\, \sigma_i) \\ &= \frac{\exp(-\psi_\Theta(\boldsymbol{x}, \boldsymbol{h}))}{(2\pi)^{n_d/2} \prod_i \sigma_i \prod_j \left(1 + \exp(d_j)\right)} \end{aligned} \tag{4}$$

where $\Theta = \{\boldsymbol{W}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{d}\}$, and

$$\begin{aligned} \psi_\Theta(\boldsymbol{x}, \boldsymbol{h}) = \sum_i \frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_i \frac{x_i - b_i}{\sigma_i^2}\boldsymbol{w}_i^T\boldsymbol{h} \\ + \sum_i \frac{1}{2\sigma_i^2}(\boldsymbol{w}_i^T\boldsymbol{h})^2 - \boldsymbol{d}^T\boldsymbol{h} \,, \end{aligned} \tag{5}$$

Compared with the Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM) (Hinton and Salakhutdinov 2006), the LRBN adopts directed links between visible nodes and hidden nodes instead of undirected links. This leads to an extra term $\sum_i \frac{1}{2\sigma_i^2}(\boldsymbol{w_i h})^2$ in Eq. (5) compared with the energy function of GRBM. This extra term explicitly captures the relationship among latent variables. The dependencies among the latent layer given the visible layer can help better explain the patterns in the input data. Furthermore, unlike the GRBM, the LRBN has no intractable partition function issue, because the joint distribution is obtained by multiplying all the prior probabilities and conditional probabilities.

Capturing spatial patterns through model learning of LRBN

We construct two LRBN models using posed and spontaneous data respectively. The input of the models are the displacements of facial feature points. Through model learning, the learned LRBN can capture both dependencies among visible variables, i.e., feature point displacements, and the dependencies among hidden variables. Thus, it can represent the feature point displacements faithfully, and capture the spatial patterns embedded in posed or spontaneous expressions successfully.

Consider the model defined in the above subsection, the goal of parameter learning is to estimate the parameters Θ given a set of data samples $\mathcal{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$ by maximizing the marginal log-likelihood,

$$\begin{aligned}\mathcal{L}(\mathcal{D};\Theta) &= \sum_m \log P_\Theta(\boldsymbol{x}^{(m)}) \\ &= \sum_m \log\left(\sum_{\boldsymbol{h}} P_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})\right) .\end{aligned} \tag{6}$$

This objective function can be maximized through gradient ascent. The exact gradient with respect to a parameter θ is,

$$\triangledown_\theta \mathcal{L}(\mathcal{D};\Theta) = \sum_m \sum_h P_\Theta(\boldsymbol{h}|\boldsymbol{x}^{(m)}) \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})}{\partial\theta} . \tag{7}$$

Computing the gradient has two difficulties:

1. Computing the posterior probability $P_\Theta(\boldsymbol{h}|\boldsymbol{x})$ is intractable even for one configuration $\boldsymbol{h}$, also known as the intractable inference;
2. There are exponentially terms to evaluate in the summation.

To address the first issue, typically variational inference algorithms are employed to approximate the true posterior distribution $P_\Theta(\boldsymbol{h}|\boldsymbol{x})$ with a factorized distribution $Q_\Phi(\boldsymbol{h}|\boldsymbol{x})$, by minimizing their KL-divergence,

$$KL(Q_\Phi(\boldsymbol{h}|\boldsymbol{x})\|P_\Theta(\boldsymbol{h}|\boldsymbol{x})) . \tag{8}$$

Some examples are the mean field algorithm (Saul, Jaakkola, and Jordan 1996), the wake-sleep algorithm (Hinton et al. 1995), and inference networks (Mnih and Gregor 2014; Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014; Gregor, Mnih, and Wierstra 2014). However, such approximations introduce a gap between the true posterior and the approximate ones, since the dependencies are not captured in the approximate distribution. In this work, we intend to preserve such dependencies by directly using the true posterior probability. Specifically, we employ Gibbs sampling to draw samples for the latent variables. One latent variable are sampled conditioned on all the other variables. Therefore, dependencies are preserved to some degree.

To address the second issue, typically Markov Chain Monte Carlo (MCMC) methods are used to estimate the summation using samples. An intuitive estimation is,

$$\triangledown_\theta \mathcal{L}(\mathcal{D};\Theta) \approx \frac{1}{n}\sum_m \sum_s \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m,s)})}{\partial\theta} , \tag{9}$$

where $\boldsymbol{h}^{(m,1)},...,\boldsymbol{h}^{(m,n)}$ are n samples from $P(\boldsymbol{h}|\boldsymbol{x}^{(m)})$. In this work, we employ the stochastic approximation procedure (SAP) framework (Robbins and Monro 1951), in which only one sample of the latent variables are used to estimate the gradient, so multiple Gibbs chains are avoided.

Under some mild assumptions, the SAP is guaranteed to converge to a local optimum (Yuille 2006) if the learning rate γ_t satisfies,

$$\begin{aligned}\sum_{t=1}^{\infty}\gamma_t &= \infty , \\ \sum_{t=1}^{\infty}\gamma_t^2 &< \infty .\end{aligned} \tag{10}$$

The gradient is then estimated as,

$$\triangledown_\theta \mathcal{L}(\mathcal{D};\Theta) \approx \sum_m \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial\theta} , \tag{11}$$

The derivative has a simple formulation because the energy function is merely a linear function of the parameters,

$$\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial w_{ij}} = \frac{h_j^{(m)}(x_i^{(m)} - \boldsymbol{w}_i^T\boldsymbol{h}^{(m)})}{\sigma_i^2} . \tag{12}$$

The gradient of other parameters can be derived similarly.

To preserve the dependencies among latent variables, we draw a sample from $P(\boldsymbol{h}|\boldsymbol{x})$ through Gibbs sampling, in which one latent node is sampled with all others fixed,

$$h_j^t \sim P(h_j|\boldsymbol{x},\boldsymbol{h}_{-j}^{t-1}) . \tag{13}$$

where $\boldsymbol{h}_{-j}$ denotes the set of all latent variables except h_j. The procedure is repeated for several iterations until mixing, and then a sample is collected for updating the parameters. Due to the requirement of taking Gibbs samples for the latent variables, the training complexity is n_h times greater than that of an RBM, if the same epochs are applied to both models.

To speed up the learning phase and scale up to large databases, we employ the stochastic gradient ascent algorithm, which estimates the gradient using a minibatch of training samples. Several passes is made over the training set until convergence. In practice, LRBN usually needs fewer epochs to converge than RBMs do. In Algorithm 1 we present the SAP for learning an LRBN.

Algorithm 1 Parameter Learning for an LRBN.

```
Input database $\mathcal{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$;
Output parameters $\Theta = \{\boldsymbol{W}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{d}\}$.
Randomly initialize the parameters $\Theta$;
Generate Gibbs samples at time step 0;
while parameters not converged, do
    Randomly choose a batch of data samples $\boldsymbol{x}$;
    Perform Gibbs sampling to obtain one sample of the latent variables for one input data, $\boldsymbol{h}^{(t)} \sim P(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{h}^{(t-1)})$;
    Compute the gradient using Eq. 12;
    Update the parameters,
    $\theta_t = \theta_{t-1} + \gamma_t \nabla_\theta \mathcal{L}(\boldsymbol{x})$ .
end while
```

Posed and spontaneous expression recognition through LRBN Inference

After training two models $\mathcal{M}_1$ and $\mathcal{M}_2$ to represent the posed and spontaneous expression respectively, given the features of a test image $\boldsymbol{x}$, binary classification is performed based on its likelihoods on the two models $P(\boldsymbol{x}|\mathcal{M}_i)$.

Directly computing $P(\boldsymbol{x})$ is intractable due to the exponentially many terms in the summation,

$$P(\boldsymbol{x}) = \sum_{\boldsymbol{h}} P(\boldsymbol{x}, \boldsymbol{h}) \,. \tag{14}$$

In this work, we estimate the log-probability using the conservative sampling-based log-likelihood (CSL) method (Bengio, Yao, and Cho 2014). The CSL estimator is based on a collection samples draws from the model given the input variables,

$$\log \hat{P}(\boldsymbol{x}) = \log \mathrm{mean}_{\boldsymbol{h} \in S} P(\boldsymbol{x}|\boldsymbol{h}) \,, \tag{15}$$

where S is a set of samples h of the latent variables collected from $P(\boldsymbol{h}|\boldsymbol{x})$. It is shown (Bengio, Yao, and Cho 2014) that the CSL estimator approaches to the ground truth log-likelihood as the length of the Markov chain approaches to infinity, and the expectation of the estimator is a lower bound of the true log-likelihood.

Experiments

Experimental conditions

As far as we know, there are several databases that include both posed and spontaneous expressions, such as the BBC smile database, the UvA-Nemo smile database (Dibeklioğlu, Salah, and Gevers 2012), the MAHNOB-Laughter database (Petridis, Martinez, and Pantic 2012), the spontaneous vs. posed facial expression (SPOS) database (Pfister et al. 2011b) and the USTC-NVIE (NVIE) database (Wang et al. 2010). The first four databases only contain the smile expression, while the last two contain six basic expression categories (i.e. happiness, sadness, anger, surprise, fear and disgust). We want to conduct experiments in the general case instead of in the case of being given a specific expression, thus we adopted the SPOS database and the NVIE database in our experiments.

Table 1: Experimental results on SPOS and NVIE databases

		SPOS database		NVIE database	
		Posed	Spontaneous	Posed	Spontaneous
Confusion matrix	Posed	49	35	501	13
	Spontaneous	21	129	0	514
Accuracy(%)	76.07			98.74	
F1 score	0.64			0.98	

The SPOS database captures posed and spontaneous expressions from 7 subject. The SPOS database consists of 84 posed and 147 spontaneous expressions, and each expression sequence starts with a onset frame and ends with a apex frame. In our experiments, the onset and apex frames of all the expression sequences are used.

The NVIE database provides both posed and spontaneous expressions. The onset and apex frames are provided for both posed and spontaneous subsets in the database. Both apex and onset frames from all posed and spontaneous expression samples, which come in pairs from the same subject are selected. During this procedure, we discarded spontaneous samples whose maximum evaluation value on six expression categories are zero, since these samples have no expression. Finally 1028 samples, including 514 posed and 514 spontaneous expression samples from 55 male and 25 female subjects, are selected.

Following Wang *et al.* (Wang, Wu, and Ji 2016), the displacement of 27 feature points (as shown in Fig. 1) between apex frames and onset frames are used as the features. Then, the features are normalized using the Z-score normalization (Abdi and Williams 2010), which makes features satisfy standard Gauss distribution and become unit-free. For the SPOS database, we adopt leave-one-subject-out cross-validation. For the NVIE database, we divide subjects into 10 groups, and each group contains 8 subjects, then we apply leave-one-group-out cross validation. During training phase, we limit the number of lattent nodes to avoid complex networks which may cause overfitting. During test phase, for each test sample, 100 thousand samples are collected from a Markov chain to approximate the exact value of log-likelihood. Accuracy and F1-score (Powers 2011) are adopted as evaluation metrics.

Experimental results and analysis

Results and analysis of posed and spontaneous expression distinction The experimental results on posed and spontaneous expression distinction are shown in Table 1. From Table 1, we can find that the proposed LRBN model achieves good performance of posed and spontaneous expression recognition on both the SPOS database and the NVIE database, with accuracy of 76.07% and 98.94%, as well as 0.64 and 0.99 of F1 score respectively. It demonstrates the effectiveness of LRBN models in capturing facial spatial patterns for posed and spontaneous expression recognition.

Comparing the results on two databases, we find that the results on the NVIE database are much better than those on the SPOS database. Since the number of samples of the NVIE database is nearly 5 times more than that of the SPOS database, and learning a LRBN model usually requires enough training data, it is reasonable that the proposed method works better on the NVIE database.

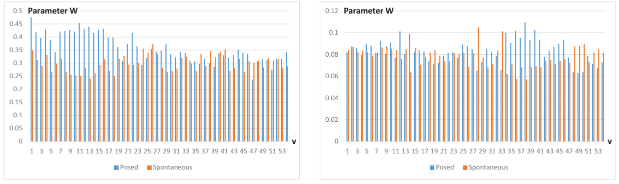
Analysis of spatial patterns inherent in feature points To investigate the spatial patterns embedded in feature point displacements for posed and spontaneous expressions, we conducted t-test to analyze whether there exist significant differences between the displacements of posed expressions and those of spontaneous expressions. The p-values are shown in Table 2. P-value less than 0.05 means significant difference. The results show that there exist more statistically significant differences between posed and spontaneous expressions on the NVIE database than those on the SPOS database. It also suggests there exists database bias. In addition, the displacements on the Y axis are significantly larger than those on the X axis for both databases. It is reasonable since the muscle movements on the Y axis are more obviously than those on the X axis.

We further analyze the spatial patterns embedded in feature point displacements for each kind of posed and spontaneous expressions (as shown in the supplementary), and find that the position with significantly differences between posed and spontaneous expressions has its own pattern for each expression. For example, for anger and surprise, there exists very few points that have significantly differences between posed and spontaneous expressions on X axis while existing much more points with significantly differences on Y axis; For sadness expression on the SPOS database, feature points 20-27 show significant differences between posed and spontaneous expression. This is consistent with the observation in (Namba et al. 2016), i.e., AU25 is one of the most frequently observed AUs in spontaneous expression but not in posed ones. Furthermore, on both databases, four expressions, i.e. disgust, fear, surprise and sadness, show greater displacements for posed expressions. On average, posed expressions also show greater displacements for all of the six expressions.

Analysis of spatial patterns captured by LRBN In order to know the mechanism of our LRBN model for capturing spatial patterns embedded in posed and spontaneous expressions, we visualize the parameter W in Fig. 2. W_i is the set of parameters w_{ij} which ties the visible node v_i and the latent node h_j's. Intuitively, the greater the parameter W_i is, the more v_i affects the captured spatial patterns. From Fig. 2, we can find that the W learning from posed expressions are quite different with that learning from spontaneous expressions on both databases, further suggesting that the spatial patterns embedded in posed and spontaneous expressions are different. Furthermore, we analyze the captured spatial patterns inherent in posed and spontaneous expression for each kind of expressions. As shown in Fig. 3 and Fig. 4, we take the captured spatial patterns of disgust and sadness from the SPOS databases for example. From Fig. 3, we can find the weights of feature points 20-24 of spontaneous disgust are larger than those of posed one, demonstrating the more importance of the facial area for spontaneous expressions. This is consistent with the observation that AU10 (i.e. the upper lip raiser, corresponds to the feature points 20-24) occurs more often in spontaneous disgust than in posed one, and AU10 is one of the three most frequently observed AUs in spontaneous disgust (Namba et al. 2016) . From Fig. 4(a), we can find that the weights of the points 1-6 in the Y axis are greater than most of the weights of the other points, demonstrating that points 1-6 play important roles in posed sadness. Since the feature points 1-6 correspond to the area of eyebrows, and these points mainly move along the Y axis, their weights in the X axis are not obvious. This further confirms that AU4, corresponding to feature points 1-6, is one of the three most frequently observed AUs in posed sadness (Namba et al. 2016) .

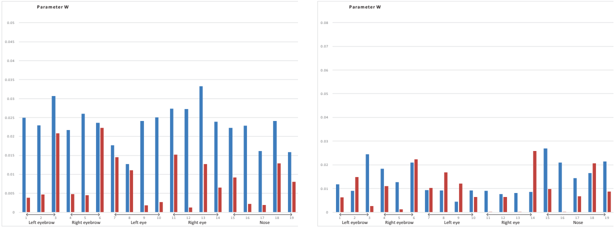
Table 2: P-values of difference between posed and spontaneous features on SPOS and NVIE databases

	SPOS database		NVIE database	
	X	Y	X	Y
1	0.080	**0.001**	**0.015**	**1.71e-13**
2	0.125	**4.86e-6**	**0.021**	**5.73e-10**
3	**0.050**	**1.58e-6**	**0.025**	**2.48e-24**
4	0.065	**1.12e-5**	0.080	**8.36e-15**
5	0.124	**5.55e-5**	**0.006**	**2.29e-6**
6	0.162	**0.005**	**0.006**	**1.02e-7**
7	0.069	0.08	0.919	**7.92e-17**
8	0.071	**0.017**	0.428	**0.046**
9	**0.038**	0.079	0.906	**1.61e-22**
10	0.062	0.029	0.677	0.309
11	0.136	**0.007**	0.276	**1.91e-22**
12	0.192	**0.028**	**7.82e-11**	**0.005**
13	0.324	**0.055**	0.893	**2.01e-5**
14	0.226	0.061	**0.024**	**0.001**
15	0.061	0.038	**0.020**	0.470
16	**0.038**	**0.113**	0.378	0.803
17	0.113	0.082	**0.001**	**0.046**
18	0.082	0.082	0.407	**2.03e-14**
19	0.199	0.199	**1.16e-11**	**6.27e-18**
20	0.768	0.204	**0.005**	**6.61e-8**
21	0.173	0.436	0.973	**0**
22	0.123	0.396	0.467	**0.002**
23	0.217	0.299	0.795	**0**
24	0.740	0.446	0.785	**7.17e-9**
25	0.148	**0.005**	**0.191**	0.084
26	0.116	**0.002**	**0.357**	**0.001**
27	0.555	**0.006**	0.251	0.069



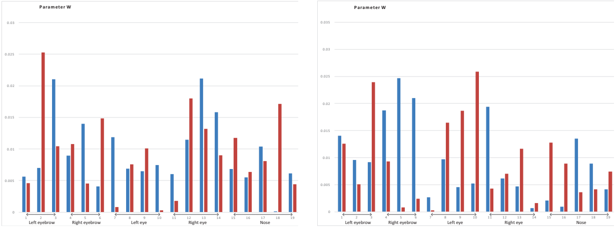
(a) (b)

Figure 2: (a) The parameter $W_i^{\prime}s$ of the model trained on SPOS database; (b) The parameter $W_i^{\prime}s$ of the model trained on NVIE database



(a) (b)

Figure 3: (a) The parameter $W_i's$ of disgust expressions along the X axis; (b) The parameter $W_i's$ of disgust expressions along the Y axis.



(a) (b)

Figure 4: (a) The parameter $W_i's$ of sadness expressions along the X axis; (b) The parameter $W_i's$ of sadness expressions along the Y axis.

Comparison with related work

We compare our work with four related works, i.e. Zhang *et al.*'s (Zhang, Tjondronegoro, and Chandran 2011), Pfister *et al.*'s (Pfister et al. 2011b), Wang *et al.*'s (Wang, Wu, and Ji 2016) and Wang *et al.*'s (Wang et al. 2015). All the related works conducted experiments on either the SPOS database or the NVIE database, or both databases.

The first two works are based on the feature-driven method. Zhang *et al.* differentiated between posed and spontaneous expressions using the SIFT appearance based features and FAP features. When they conducted experiments on the NVIE database, they selected 3572 posed and 1472 spontaneous images as the samples. Since Zhang did not provide enough details about how they selected samples, we can not obtain the same samples as theirs. We just compare the experimental results for reference. Pfister *et al.* proposed the spatio-temporal local texture descriptor and use it for posed and spontaneous expression differentiation. They conducted experiments on the SPOS database.

The last two works employ model-based approaches. Wang *et al.* (Wang et al. 2015) proposed multiple Bayesian networks to capture posed and spontaneous spatial facial patterns from feature points respectively. Wang *et al.* (Wang, Wu, and Ji 2016) proposed to use restricted Boltzmann machines to explicitly model complex joint distributions over feature points.

The comparison results are shown in Table 3. From Table 3, we can obtain the following observations:

Table 3: Comparison with related works

		Accuracy(%)	F1 score
SPOS database	L. Zhang *et al.* (Zhang, Tjondronegoro, and Chandran 2011)	72.00	/
	Wang *et al.* (Wang et al. 2015)	74.79	**0.67**
	Wu *et al.* (Wang, Wu, and Ji 2016)	76.07	0.64
	Ours	**76.07**	0.64
NVIE database	Pfister *et al.* (Pfister et al. 2011b)	79.43	/
	Wang *et al.* (Wang et al. 2015)	91.63	0.91
	Wu *et al.* (Wang, Wu, and Ji 2016)	92.61	0.92
	Ours	**98.74**	**0.98**

First, the three model-based approaches outperform the two feature-driven approaches on both databases. The two feature-driven works use appearance features and geometric features, while the three model-based approaches adopt geometric features only. Although two feature-driven works extracted more complex features, model-based approaches achieve better performance. This demonstrates that the model-based approaches can successfully capture spatial patterns embedded in posed and spontaneous expressions, and effectively leverage such spatial patterns for posed and spontaneous expression distinction.

Second, among the three model-based approaches, our proposed LRBN model achieves the best performance with highest evaluation metrics in most cases. Specifically, We achieve the same performance as Wang *et al.*'s (Wang, Wu, and Ji 2016), and better performance than Wang *et al.*'s (Wang et al. 2015) on the SPOS database, and much better performance on the NVIE database. Unlike BN, which can only model local dependencies among variables, the proposed LRBN can capture global dependencies among variables through introducing hidden units. Furthermore, unlike the RBM, which capture global dependencies among visible units through introducing hidden units, but hidden units are independent to each other given visible units, the proposed LRBN can capture the dependencies among the latent variables given the observation. These dependencies are crucial for faithful data representation, demonstrated by the superior performance of the LRBN.

Conclusion

In this paper, we propose LRBNs to explicitly model complex joint distributions over feature points, i.e., spatial patterns, embedded in posed and spontaneous expressions respectively, and leverage such spatial patterns for posed and spontaneous expression distinction. Specifically, we construct two LRBNs to model spatial patterns embedded in posed and spontaneous expressions respectively. During training, an efficient algorithm is proposed to learn the parameters of LRBNs. During testing, the samples are classified into posed or spontaneous expressions according to the LRBN with the largest likelihood, which is estimated by CSL method. Experimental results on two benchmark databases demonstrate the power of the proposed model in capturing spatial patterns as well as its advantage over existing methodologies for posed and spontaneous expression distinction.

Acknowledgement

This work has been supported by the National Science Foundation of China (Grant No. 61175037, 61228304,61473270), and the project from Anhui Science and Technology Agency (1508085SMF223).

References

Abdi, H., and Williams, L. 2010. Normalizing data. *Encyclopedia of research design* 935–938.

Bengio, Y.; Yao, L.; and Cho, K. 2014. Bounding the test log-likelihood of generative models. In *International Conference on Learning Representations (Conference Track)*.

Cohn, J., and Schmidt, K. 2004. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing* 2(02):121–132.

Dibeklioglu, H.; Valenti, R.; Salah, A. A.; and Gevers, T. 2010. Eyes do not lie: spontaneous versus posed smiles. In *Proceedings of the international conference on Multimedia*, 703–706. ACM.

Dibeklioğlu, H.; Salah, A.; and Gevers, T. 2012. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *ECCV 2012*. Springer. 525–538.

Dibeklioglu, H.; Salah, A.; and Gevers, T. 2015. Recognition of genuine smiles. *Multimedia, IEEE Transactions on* PP(99):1–1.

Ekman, P., and Friesen, W. 1982. Felt, false, and miserable smiles. *Journal of nonverbal behavior* 6(4):238–252.

Ekman, P.; Hager, J.; and F., W. 1981. The symmetry of emotional and deliberate facial actions. *Psychophysiology* 18(2):101–106.

Gan, Q.; Wu, C.; Wang, S.; and Ji, Q. 2015. Posed and spontaneous facial expression differentiation using deep boltzmann machines. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, 643–648. IEEE.

Gregor, K.; Mnih, A.; and Wierstra, D. 2014. Deep autoregressive networks. *In Proceedings of the 31st International Conference on Machine Learning*.

Hinton, G., and Salakhutdinov, R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Hinton, G. E.; Dayan, P.; Frey, B. J.; and Neal, R. M. 1995. The” wake-sleep” algorithm for unsupervised neural networks. *Science* 268(5214):1158–1161.

Hinton, G. 2010. A practical guide to training restricted boltzmann machines. *Momentum* 9(1):926.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Littlewort, G.; Bartlett, M.; and Lee, K. 2009. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing* 27(12):1797–1803.

Mnih, A., and Gregor, K. 2014. Neural variational inference and learning in belief networks. *In Proceedings of the 31st International Conference on Machine Learning*.

Namba, S.; Makihara, S.; Kabir, R. S.; Miyatani, M.; and Nakao, T. 2016. Spontaneous facial expressions are different from posed facial expressions: Morphological properties and dynamic sequences. *Current Psychology* 1–13.

Petridis, S.; Martinez, B.; and Pantic, M. 2012. The mahnob laughter database. *Image and Vision Computing*.

Pfister, T.; Li, X.; Zhao, G.; and Pietikainen, M. 2011a. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *ICCV Workshops*, 868–875. IEEE.

Pfister, T.; Li, X.; Zhao, G.; and Pietikäinen, M. 2011b. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 868–875. IEEE.

Powers, D. M. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1278–1286.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Saul, L. K.; Jaakkola, T.; and Jordan, M. I. 1996. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research* 4(61):76.

Seckington, M. 2011. Using dynamic bayesian networks for posed versus spontaneous facial expression recognition. *Mater Thesis, Department of Computer Science, Delft University of Technology*.

Valstar, M.; Pantic, M.; Ambadar, Z.; and Cohn, J. 2006. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, 162–170. ACM.

Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; and Wang, X. 2010. A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on* 12(7):682–691.

Wang, S.; Wu, C.; He, M.; Wang, J.; and Ji, Q. 2015. Posed and spontaneous expression recognition through modeling their spatial patterns. *Machine Vision and Applications* 1–13.

Wang, S.; Wu, C.; and Ji, Q. 2016. Capturing global spatial patterns for distinguishing posed and spontaneous expressions. *Computer Vision and Image Understanding* 147:69–76.

Wu, C., and Wang, S. 2016. Posed and spontaneous expression recognition through restricted boltzmann machine. In *MultiMedia Modeling*, 127–137. Springer.

Yuille, A. L. 2006. The convergence of contrastive divergences. *Department of Statistics, UCLA*.

Zhang, L.; Tjondronegoro, D.; and Chandran, V. 2011. Geometry vs. appearance for discriminating between posed and spontaneous emotions. In *Neural Information Processing*, 431–440. Springer.