Capturing Spatial and Temporal Patterns for Distinguishing between Posed and Spontaneous Expressions

Jiajia Yang
University of Science and Technology of China
Hefei, Anhui
yang25@mail.ustc.edu.cn

Shangfei Wang*
University of Science and Technology of China
Hefei, Anhui, P.R.China
sfwang@ustc.edu.cn

ABSTRACT

Spatial and temporal patterns inherent in facial behavior carry crucial information for posed and spontaneous expressions distinction, but have not been thoroughly exploited yet. To address this issue, we propose a novel dynamic model, termed as interval temporal restricted Boltzmann machine (IT-RBM), to jointly capture global spatial patterns and complex temporal patterns embedded in posed and spontaneous expressions respectively for distinguishing between posed and spontaneous expressions. Specifically, we consider a facial expression as a complex activity that consists of temporally overlapping or sequential primitive facial events, which are defined as the motion of feature points. We propose using the Allen's Interval Algebra to represent the complex temporal patterns existing in facial events through a two-layer Bayesian network. Furthermore, we propose employing multi-value restricted Boltzmann machine to capture intrinsic global spatial patterns among facial events. Experimental results on three benchmark databases, the UvA-NEMO smile database, the DISFA+ database, and theSPOS database, demonstrate the proposed interval temporal restricted Boltzmann machine can successfully capture the intrinsic spatial-temporal patterns in facial behavior, and thus outperform state-of-the art work of posed and spontaneous expressions distinction.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Computer vision tasks**;

KEYWORDS

Global spatial and temporal patterns, Posed and spontaneous expressions distinction, Multi-value RBM, Interval temporal restricted Boltzmann machine(IT-RBM), Allen's Interval Algebra

*The corresponding author.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123350

1 INTRODUCTION

Posed and spontaneous expression distinction has attracted increasing attention in recent years due to its wide potential application in human-computer interaction. Posed expressions may deliberately disguise inner feelings, while spontaneous expressions convey the true emotions. Since expressions are affected by many significant subject dependent variations, and the facial appearance differences between posed and spontaneous expressions are very subtle, distinguishing posed and spontaneous facial expressions is rather challenging. Successfully capturing the inherent spatial and temporal patterns may facilitate posed and spontaneous expression distinction.

Behavior research has indicated that there exist differences in temporal and spatial patterns between posed and spontaneous expressions. Temporal patterns involve the speed, amplitude, trajectory and total duration of onset and offset. For example, Ekman *et al.* [7][6] revealed that compared with posed expressions, the trajectory appears often smoother, the total duration is usually shorter, and onset is more gradual for spontaneous expressions in most cases.

Schmidt *et al.* [18] proved that maximum speed of movement onset is greater in deliberate smiles than spontaneous smiles. Maximum speed and amplitude are greater and duration is shorter in deliberate eyebrow raises compared to spontaneous eyebrow raises. Spatial patterns mainly consist of the movement of facial muscles. Ekman *et al.*'s work [7] indicated that orbicularis oculi are contracted during spontaneous smiles, but not during posed smiles; Although the zygomatic major is contracted for both posed and spontaneous expressions, its contraction is more likely to occur asymmetrically for posed smiles than spontaneous smiles [8]. Ross and Pulusu [16] proved that posed expressions begin on the right face whereas spontaneous expressions begin on the left face in most cases, especially for upper facial expressions, such as smile and surprise.

Inspired by the observations from behavior research, most work defines specific features for posed and spontaneous expression distinction. For example, Cohn and Schmidt [2] adopted temporal features, such as duration, amplitude, and the ratio of amplitude to duration. Valstar [20] extracted speed, intensity, duration, symmetry, trajectory and the occurrence order of brow actions from the displacements of facial fiducial points. Dibeklioglu *et al.* [4] extracted amplitude, duration, speed, and acceleration to describe dynamics of eyelid, cheek, and lip corner movements. Seckington

[19] defined six features to represent temporal dynamics, including morphology, apex overlap, symmetry, total duration, speed of onset and speed of offset.

After feature extraction, both static classifiers and dynamic classifiers were investigated to distinguish posed and spontaneous expressions. The static classifiers, such as the linear discriminant classifier [2], support vector machine [12] , k-NN and naive Bayesian classifiers [5], were used to model the mapping between features and expression types. The dynamic classifiers, such as continuous hidden Markov model [5] and dynamic Bayesian network [19] were adopted to model the temporal dynamic for posed and spontaneous expression distinction. All these research demonstrated the progress in distinguishing posed and spontaneous expressions. However, most current work employed different features and classifiers for posed and spontaneous expression distinction, without explicitly capturing spatial and temporal patterns embedded in posed and spontaneous expression respectively, and leverage such spatial and temporal patterns for posed and spontaneous expression distinction. We call them feature-driven methods.

Only recently, Wang *et al* [21] proposed multiple Bayesian networks (BN) to capture spatial patterns for posed and spontaneous expressions respectively given gender and expression categories. We call it a model-based method. However, due to the first-order Markov assumption of BN, their model can only capture the local dependencies among geometric features instead of the global and high-order relations among them. To address this issue, Wu *et al* [22] proposed to use restricted Boltzmann machine to explicitly model complex joint distributions over feature points , i.e. spatial patterns, embedded in posed and spontaneous expressions respectively, since RBM can model higher-order dependencies among random variables by introducing a layer of latent units [10]. 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. Therefore, Quan *et al.* [9] proposed to employ latent regression Bayesian network to capture the high-order and global dependencies among facial geometric features. Unlike RBM, which is an undirected latent variable model, latent regression Bayesian network is a directed latent variable 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. Such dependencies are crucial for faithful data representation. All the three model-based work demonstrate that explicitly capturing spatial patterns are helpful to differentiate between posed and spontaneous expressions.

However, to the best of our knowledge, little work explicitly capture both spatial patterns and temporal patterns embedded in posed and spontaneous expression, and leverage such spatial and temporal patterns for posed and spontaneous expression distinction. Therefore, in this paper, we introduce a novel dynamic model, termed as interval temporal restricted Boltzmann machine(IT-RBM), to jointly capture global spatial patterns and complex temporal patterns embedded in posed expressions and spontaneous expressions respectively for distinguishing posed and spontaneous expressions. The proposed IT-RBM is a three-layer hierarchical probabilistic graphical model. The top two layer is a multi-value RBM, capture global spatial patterns, and the bottom tow layer is a BN, modeling temporal patterns.

The proposed IT-RBM is a novel dynamic model. Unlike commonly used dynamic model, such as HMM and DBN, which can only handle three time point relations (precedes, follows, equals), the proposed IT-RBM can provide more complex relations through incorporating interval algebra. Instead of capturing local stationary dynamics only due to the assumption of the first order Markov property and stationary transition, like HMM and DBN, the proposed IT-RBM can model global temporal relations.

The most related dynamic model to our model is interval temporal Bayesian network (ITBN) proposed by Wang *et al.* [23]. An ITBN implemented as a BN, includes two types of nodes: temporal entity nodes and temporal relation nodes. The links connecting temporal entity nodes capture the spatial dependencies among the temporal entities. The links connecting the temporal relation nodes with the corresponding temporal entities characterize the temporal relationships between the two connected temporal entities. Thus, the ITBN can capture both spatial patterns and temporal patters. Wang et al. apply ITBN for expression recognition. Unlike ITBN, which used BN to capture both spatial and temporal relationships, we propose a hybrid graphic model, including a RBM to represent spatial patterns and a BN to represent temporal patterns. Therefore, our model can capture high order spatial patterns, while ITBN can only model local spatial patterns due to the first order Markov property of BN.

Compared with related work, the contributions of the paper is as follows:

1. We propose a novel dynamic model, i.e. IT-RBM, which can capture global spatial patterns and complex temporal patterns jointly.

2. We propose to explicitly model spatial-temporal patterns inherent in posed expressions and spontaneous expressions respectively for posed and spontaneous expression distinction.

2 PROPOSED METHOD

We consider a posed expression or spontaneous expression as a complex activity consisting of sequential or temporally overlapping primitive facial events. Similar to [23], the primitive facial events are defined as the motion of facial feature points. Therefore, the movement of one feature point corresponds to one primitive facial event, which records the motion state, the starting time when the feature point leaves its

neutral position, and the ending time when the feature point comes back to its neutral position. The interval relation between every pair of primitive facial events can be defined as one of 13 interval relations proposed by Allen's interval algebra [1]. The primitive event pairs and their interval relations which have larger variance between posed expressions and spontaneous expressions are selected for posed and spontaneous expression distinction. Then, we build two IT-RBM models with the selected primitive events and interval relations for posed expressions and spontaneous expressions respectively. During training, the constructed IT-RBMs capture the global spatial and temporal patterns jointly. During testing, the label of a test sample is the model with larger likelihood. The framework of our method is shown as Fig.1.In this section, we will introduce primitive facial events extraction and temporal relations definition and selection firstly, then present the proposed IT-RBM model exhaustively.

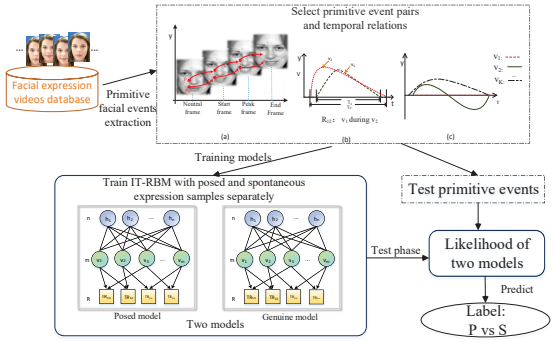

Figure 1: The outline of recognition system

2.1 Primitive facial events extraction

Given sample videos of posed and spontaneous expressions donated as $S = \{(X_\alpha, Y_\alpha), \alpha = 1, ..., Q\}$, where X_α is the α^{th} video with frame length of f_α, and Y_α is its expression label, Q is the number of total videos. Each frame is a facial image with P_{no} facial points.

We consider each expression video consists of sequential or temporally overlapping primitive facial events. The primitive facial events are defined as the motion of facial feature points. Therefore, the movement of one feature point corresponds to one primitive facial event, which records the motion state, the starting time when the feature point leaves its neutral position, and the ending time when the feature point comes back to its neutral position. A primitive event shown as Fig.2, denoted as $V = \langle ts, te, K \rangle, (ts, te \in \mathbb{R}, ts < te)$, ts and te denote the start time and the end time, respectively. K is a set of all possible states for the primitive events extracted by k-means clustering.

Since the frame length of expression videos are not the same, we normalize the frame length of all expression videos to len, which is the smallest frame length in the training set. The samples whose frame length is larger than len are

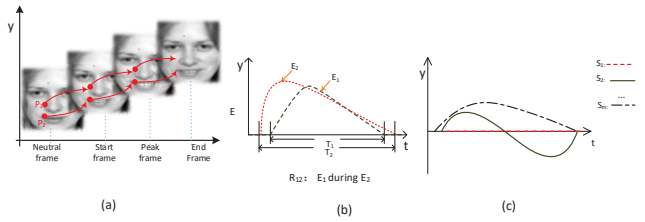

Figure 2: (a) Facial muscle movement as captured by the movement of facial points.(b) Duration for event v_1 and v_2 and their temporal relation.(c) Typical movement patterns of a primitive facial event.

equidistantly downsampled to len frames. Then we adopt K-means clustering on the feature point displacement sequence $[P_{no} \times len]^Q$ to obtain K movement models.

2.2 Temporal relations definition and selection

Allen's interval algebra [1] defines 13 temporal relations, i.e. $\mathbb{I} = \{b, bi, m, mi, o, oi, d, di, f, fi, eq\}$, respectively representing before, meets, overlaps, starts, during, finishes, equals and their inverses. Through calculating temporal distance $dis(v_i, v_j)$ according to Eq. 1, we can obtain the temporal relation between each pair of primitive facial events from Table. 1.

$$dis(v_i, v_j) = [ts_i - ts_j, ts_i - te_j, te_i - ts_j, te_i - te_j]; \quad (1)$$

Table 1: TD and Interval relation mapping table

No	TR	$ts_i - ts_j$	$te_i - te_j$	$ts_i - te_j$	$te_i - ts_j$	illustration
1	b	< 0	< 0	< 0	< 0	
2	bi	> 0	> 0	> 0	> 0	
3	d	> 0	< 0	< 0	> 0	
4	di	< 0	> 0	< 0	> 0	
5	o	< 0	< 0	< 0	> 0	
6	oi	> 0	> 0	< 0	> 0	
7	m	< 0	< 0	< 0	$= 0$	
8	mi	> 0	> 0	$= 0$	> 0	
9	s	$= 0$	< 0	< 0	> 0	
10	si	$= 0$	> 0	< 0	> 0	
11	f	> 0	$= 0$	< 0	> 0	
12	fi	< 0	$= 0$	< 0	> 0	
13	eq	$= 0$	$= 0$	–	–	

After primitive facial event extraction, we obtain $P_{no} * (P_{no} - 1)$ pairs of primitive facial events and their corresponding temporal relations for every sample. The discriminative temporal relations should have larger variance between posed and spontaneous expressions, therefore, we proposed to use Kullback-Leibler divergence [15] based score to measure the difference between two probability distributions.

$$Score_{ij} = \sum D_{KL}(P_{v_i} \| P_{v_j}) + D_{KL}(P_{v_j} \| P_{v_i}) \quad (2)$$

The score of event pair v_i, v_j is defined in Eq.2, where P_{v_i} P_{v_j} is the conditional probability of relation between event pair v_i, v_j takes all value $TR \in \mathbb{I}$ for the $v_i^{th}(v_j^{th})$ expression with $v_i(v_j)$ ranging over all expressions, v_i and v_j represent posed and spontaneous expression singly. D_{KL} stands for the KL divergence. All the primitive event pairs are ranked according to their score. The top π pairs with m primitive events are selected and their temporal relations will be instantiated in different samples and temporal relations are denotes by the bottom layer nodes of IT-RBM model.

For example, the upside of Fig. 1 illustrates primitive facial events. Specifically, (a) shows two primitive facial events. Facial point P_1 and P_2 correspond to event v_1 and v_2 and represent muscle motion of left wing of nose and right mouth corner respectively. (b) simply draw the curve chart of two events and T_1, T_2 are the corresponding duration for V_1 and V_2. However, (b) is only the trace along the vertical direction. Every primitive event has K possible states, which represent their movement patterns over the time interval as shown in (c). The first red line represents the point staying still throughout the process. The other states represent $k-1$ movement patterns. For example, state v_k denoted by black dotted line represents that the point moves up and then comes back. State v_2 denoted by green solid line shows a relatively more complex pattern in which the point moves up in the beginning and moves down later.

2.3 Spatial and temporal patterns capturing through IT-RBM Model

The proposed model IT-RBM is shown in Fig. 3. It is a hybrid graphic model, whose top part is a multi-value RBM and the bottom part is a Bayesian network. From top to bottom, the first layer contains n binary latent variables $h_n \in \{0,1\}$. The second layer contains m visible nodes, $V_m^k \in \{1,...,K\}$ represents m selected primitive facial events with K motion states, and the bottom layer contains R temporal relation nodes, $TR \in \{1,...,13\}$ represent 13 temporal relations. The bottom part captures the complex temporal relations while the top part models the global dependencies among facial events, i.e. spatial patterns inherent in posed or spontaneous expressions. The joint probability of the proposed IT-RBM is shown in Eq. 3

$$P(V,TR) = \sum_H P(V,TR,H) = P(TR|V) \sum_H P(V,H) \quad (3)$$

where

$$P(TR|V) = \prod_{r=1}^{R} P(TR_r|\pi(TR_r)), \quad (4)$$

and TR_r represents the r^{th} temporal relation node, $\pi(TR_r)$ are the two primitive event nodes that produce TR_r. $P(V,H)$ is the joint probability of the top part multi-value RBM.

After primitive events extraction, we obtain training data $\{v_\alpha, L_\alpha\}^{Q_{train}}$, where Q_{train} indicates the number of training samples of spontaneous or posed expressions. The goal of

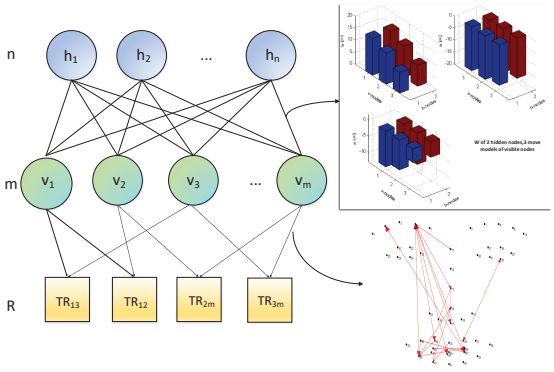

Figure 3: An example of IT-RBM model

model learning is to maximize the log likelihood as follows:

$$\theta^* = \arg\max_\theta \left(\frac{1}{Q_{train}} \sum \log P(v|\theta) + \log P(TR|v) \right) \quad (5)$$

From Eq. 5, we can find that the log likelihood of IT-RBM can be factorized into the sum of the log likelihood of RBM and the log likelihood of BN. Therefore, we can train RBM and BN separately.

As for multi-value RBM, the marginal distribution of the visible units is calculated as Eq. 6,

$$\begin{aligned} p(v) &= \frac{1}{Z} \sum_h p(v,h) = \frac{1}{Z} \sum_h e^{-E(v,h)} \\ &= \frac{1}{Z} e^{\sum_{i=1}^{m} \sum_{k=1}^{K} b_i^k v_i^k} \prod_{j=1}^{n} (1 + e^{a_j + \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ij}^k v_i^k}) \end{aligned} \quad (6)$$

where E is the energy function of multi-value RBM and is defined in Eq. 7. $\{W, a, b\}$ are the model parameters: w_{ij}^k is a symmetric interaction term between visible unit i that takes on value k, and hidden unit$j$, b_i^k is the bias of unit i that takes on value k, and a_j is the bias of hidden unit j.

$$E(v,h) = -\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{K} v_i^k w_{ij}^k h_j - \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k b_i^k - \sum_{j=1}^{n} h_j a_j \quad (7)$$

The gradient with respect to $\Theta_R = \{w, a, b\}$ can be calculated as Eq. 8, where the angle brackets are used to denote expectations under the distribution specified by the subscript that follows, and ε is a learning rate.

$$\Delta\Theta_R = \varepsilon \frac{\partial \log p(v)}{\partial \Theta_R} = \varepsilon \left(\left\langle \frac{\partial E}{\partial \Theta_R} \right\rangle_{data} - \left\langle \frac{\partial E}{\partial \Theta_R} \right\rangle_{model} \right) \quad (8)$$

Calculating the gradient involves inferring $P(h,v|\Theta)$, which is intractable, so we use CD algorithm [11]. The conditional distributions are given by softmax and logistic functions as follows:

$$p(v_i^k = 1|h) = \frac{\exp(b_i^k + \sum_{j=1}^{n} h_j w_{ij}^k)}{\sum_{l=1}^{K} \exp(b_i^l + \sum_{j=1}^{n} h_j w_{ij}^l)} \quad (9)$$

$$p(h_j = 1|v) = \sigma(a_j + \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ij}^k v_i^k) \quad (10)$$

The algorithm of leaning multi-value RBM's parameters using CD learning is shown as Algorithm 1.

Algorithm 1 The training algorithm for multi-value RBM using CD learning

Require: Training data:$\mathbf{v}_\alpha = \{0, 1, ..., K\}^m$, latent nodes number $\mathbf{n}$, learning rate $\boldsymbol{\epsilon}$, maximum training time $\mathbf{T}$
Ensure: $\mathbf{W_{ij}^k}, \mathbf{a_j}, \mathbf{b_i^k}$

```
Initialize: set W, a, b to small random values
for t = 1, 2, ..., T do
    for j = 1, 2, ..., n do
        Sample h_αj ~ p(h_αj = 1|v_α) with Eq.10
    end for
    for i = 1, 2, ..., m do
        Sample v^k_βi ~ p(v^k_βi = 1|h_α) with Eq.9
    end for
    parameters update:
    W ← W + ε(P(h_α = 1|v_α)v_α^T − P(h_β = 1|v_β)v_β^T)
    b ← b + ε(v_1 − v_2)
    a ← a + ε(P(h_α = 1|v_α) − P(h_β = 1|v_β))
end for
```

As for BN part, parameters for BN involve the conditional probability distribution for each temporal relation node TR_{ij} given its parents nodes v_i and v_j. After selection routine, BN structure and the number of parameters are determined. The conditional probability of temporal relation TR_{ij} given the primitive event pair v_i v_j holds regardless of their moving patterns, only involved with their start and end time. Obtained training data $\{v_\alpha, L_\alpha\}^{Q_{train}}$ contains the interval time of each primitive event and their temporal relations, the goal of parameter estimation is to find the maximum likelihood estimate of parameter Θ_B, shown as Eq. 11.

$$\Theta_B = \arg\max_{\Theta_B} \log \sum_{Q_{train}} \log P(TR|v; \Theta_B) \tag{11}$$

2.4 Posed and spontaneous expression distinction

After training, we obtain an IT-RBM model given posed and spontaneous expressions separately. During testing, only geometric features provided. The label of a test sample t is the class with greater log likelihood value according to Eq. 12. Where l^* is the predicted label, C is the number of IT-RBM models.

$$l^* = \max_{l \in [1,C]} \{\text{logP}\,(t|\theta_l)\} \tag{12}$$

For the test sample t, the log likelihood that IT-RBM trained on class l assign to t is as follows:

$$logP(t|\theta_l) = \log\left(\sum_h \exp(-E(h, t; \theta_l))\right) - \log Z\,(\theta_l) + \log(P(TR|t; \theta_l)) \tag{13}$$

where, it is intractable to calculate partition function Z directly. Salakhutdinov and Murry proposed to use Annealed Importance Sampling (AIS) [17] to estimate the partition function of an RBM with binary visible units. Inspired by this, we extend the AIS method to calculate the partition function of multi-value RBM.

AIS estimates the ratio of partition function of the object RBM to a base-rate RBM. Suppose we have two multi-value RBM with parameter values $\theta_A = \{W^A, b^A, a^A\}$ and $\theta_B = \{W^B, b^B, a^B\}$ that define probability distributions p_A and p_B over same $v \in \{0, 1, ..., K\}^m$, and $h^A \in \{0, 1\}^{n_A}, h^B \in \{0, 1\}^{n_B}$.

First, a sequence of intermediate distributions for $\tau = 0, ..., \tau$ are defined as:

$$p_\tau(v) = \frac{p^*_\tau(v)}{Z_\tau} = \frac{1}{Z_\tau} \sum_h \exp(-E_\tau(v, h)), (\tau = 0, 1, ..., \tau), \tag{14}$$

where energy function is defined as Eq. 15. The unnormalized probability over visible units can be estimated as Eq. 16, where $0 = \beta_0 < \beta_1 < ... < \beta_\tau = 1$.

$$E_\tau(v, h) = (1 - \beta_\tau) E(v, h^A; \theta_A) + \beta_\tau E(v, h^B; \theta_B) \tag{15}$$

$$P^*_\tau(v) = e^{(1-\beta_\tau)\sum_i \sum_k b_i^{kA} v_i^k} \prod_{j=1}^{n_A} (1 + e^{(1-\beta_\tau)(\sum_i \sum_k W_{ij}^{kA} v_i^k + a_j^A)}) * e^{\beta_\tau \sum_i \sum_k b_i^{kB} v_i^k} \prod_{j=1}^{n_B} (1 + e^{\beta_\tau (\sum_i \sum_k W_{ij}^{kB} v_i^k + a_j^B)}) \tag{16}$$

Then, we define a Markov chain transition operator $T_\tau(v'; v)$ that leaves $p_\tau(v)$ invariant. The conditional distributions are given by softmax and logistic functions:

$$p(h_j^A = 1|v) = \sigma((1 - \beta_\tau)(a_j^A + \sum_i \sum_k \left(w^A\right)_{ij}^k v_i^k)) \tag{17}$$

$$p(h_j^B = 1|v) = \sigma(\beta_\tau(a_j^B + \sum_i \sum_k \left(w^B\right)_{ij}^k v_i^k)) \tag{18}$$

$$p(v_i^{'k} = 1|h) = (1 - \beta_\tau) \frac{\exp(\left(b^A\right)_i^k + \sum_{j=1}^n h_j^A \left(w^A\right)_{ij}^k)}{\sum_{l=1}^K \exp((b^A)_i^l + \sum_{j=1}^n h_j^A \left(w^A\right)_{ij}^l)} + \beta_\tau \frac{\exp(\left(b^B\right)_i^k + \sum_{j=1}^n h_j^B \left(w^B\right)_{ij}^k)}{\sum_{l=1}^K \exp((b^B)_i^l + \sum_{j=1}^n h_j^B \left(w^B\right)_{ij}^l)} \tag{19}$$

Eqs. 17 and 18 are used to stochastically activate hidden units h_A and h_B. Eq. 19 is then used to draw a new sample v'. Finally, initial $\theta_A = \{0, b^A, 0\}$, Z^A is calculated as Eq. 20.

$$Z^A = 2^{n_A} \prod_i \prod_k e^{b_i^k} \tag{20}$$

The detailed algorithm is shown as Algorithm 2

In Eq. 13, $log(P(I|t))$ is the likelihood of bottom BN, when given test sample t combining with selection routine their selected primitive facial event pairs and temporal relations are determined, so the probability of $log(P(I|t))$ can be calculated easily.

Algorithm 2 The AIS algorithm for capturing partition function Z[17]

Require: Base-rate RBM's parameters $\Theta_A = \theta_0$, Objective RBM's parameters $\Theta_B = \theta_\beta$, $\beta \in [0, 1]$,
Ensure: Objective RBM's Z_B
1: **for** i = 1 to M_r **do**
2: **for** $\beta = 0$ to 1 **do**
3: Generate $v_1, v_2, ..., v_\tau$ using $T_\tau$as follows:
4: Sample v_1 from $p_A = p_0$
5: Sample v_2 given v_1 usingT_1
6: ...
7: Sample v_τ given $v_{\tau-1}$ using $T_{\tau-1}$
8: **end for**
9: **end for**
10: $\frac{Z_B}{Z_A} \approx \frac{P_1^*(v_1)}{P_0^*(v_0)} \frac{P_2^*(v_2)}{P_1^*(v_1)} \cdots \frac{P_\tau^*(v_\tau)}{P_{\tau-1}^*(v_{\tau-1})} = \frac{1}{M_r} \sum_{i=1}^{M_r} \omega^{(i)} = \widehat{r}_{AIS}$
11: $Z_B = Z_A * \widehat{r}_{AIS}$

3 EXPERIMENT

3.1 Experimental Condition

In order to evaluate the performance of the proposed method, we conduct posed and spontaneous expression distinction experiments on three benchmark datasets, the UvA-NEMO Smile (UvA-NEMO) database [3], the Extended DISFA (DISFA+) dataset [13] and the SPOS dataset [14].

The UvA-NEMO database, the largest posed and spontaneous smile database, consists of 597 spontaneous smile videos and 643 posed smile videos from 400 subjects. The DISFA+ database includes 572 posed and 252 spontaneous expression videos with five expression categories, i.e., disgust, fear, happiness, sadness, and surprise of 27 young adults. The SPOS database provides visible and near infrared images, we utilize the visible images, including 84 posed and 150 spontaneous expression samples of 7 subjects with six expression categories,i.e., anger, disgust, fear, happiness, sadness and surprise. The data distribution of three databases are show in Table.2. We define facial primitive event as

Table 2: Number of samples

	SPOS	
Expression	POSED	SPON
Anger	14	13
Disgust	14	23
Fear	14	32
Happy	14	66
Sad	14	5
Surprise	14	11
Total	84	150

	DISFA+	
Expression	POSED	SPON
Disgust	163	81
Fear	163	63
Happy	42	18
Sad	122	54
Surprise	82	36
Total	572	252

UvA-NEMO	
POSED	SPON
553	473

the motion of facial feature point in Section 2.1, therefore we extract facial landmark points as features. For the UvA-NEMO database and the SPOS database, 49 facial feature points shown in the bottom right corner of Fig. 1 are extracted using SDM [26]. For the DISFA+ database, 68 facial points are provided by database constructors. We utilize 49 points without the points of face outline. Recognition accuracy and F1-score are used as performance metrics, and 10-fold subject-independent cross validation is adopted on the UvA-NEMO database and the DISFA+ database. Since the SPOS database consists of videos from only 7 subjects, 5-fold subject-independent cross validation is adopted.

We conduct posed and spontaneous expression distinction experiments with three methods. The first one is the posed IT-RBM, which captures spatial and temporal patterns simultaneously. The second one is the multi-value RBM, the top part of the proposed IT-RBM which only models high-order spatial patterns. The last one is HMM, the commonly used dynamic model which can only obtain temporal patterns.

3.2 Experimental results and analyses

Results of posed and spontaneous expression distinction experiments with three methods on three datasets are shown in Table 3. From Table 3, we can obtain the following observations:

First, the proposed IT-RBM performs better than multi-value RBM with higher accuracy and F1-score on all databases. Since the proposed IT-RBM can captures spatial and temporal patterns embedded in posed and spontaneous expression simultaneously, while the multi-value RBM only model inherent spatial patterns. Furthermore, compared with multi-value RBM, the improvement of the proposed IT-RBM is more significant on the DISFA+ database and the SPOS database than the UvA-NEMO database. Specifically on the DISFA+ database and the SPOS database, the accuracy of IT-RBM is 0.0667 and 0.0085 higher than those of multi-value RBM and F1-score is increased by 0.0555 and 0.0166 respectively. While on the UvA-NEMO database, the accuracy and F1-score both increased by 0.0032. Since the DISFA+ database and the SPOS database consist of posed and spontaneous expressions with five and six expression categories respectively, while the UvA-NEMO database includes only one expression category, i.e. posed and spontaneous smile. Therefore, compared with the UvA-NEMO Smile database, the posed and spontaneous expression distinction is more challenging on the DISFA+ database and the SPOS database with more expression categories. In such case, the extra temporal patterns captured by IT-RBM benefits the task of the posed and spontaneous expression distinction more significantly.

Second, the proposed IT-RBM is superior to HMM with higher accuracy and F1-score on all databases. As shown in Table 3, the accuracy of HMM method is lower than IT-RBM by 0.3011, 0.1092 and 0.1325, and F1-score of HMM is lower than IT-RBM by 0.2021, 0.0570 and 0.1013 on the UvA-NEMO database, the DISFA+ database and the SPOS database respectively. HMM, the commonly used dynamic model, can only handle three time point relations (precedes, follows, equals), and capture local stationary dynamics due to the assumption of the first order Markov property and stationary transition. While the proposed IT-RBM can provide

Table 3: Experiment results on three databases

dataset	UvA-NEMO			DISFA+			SPOS		
method	HMM	RBM*	IT-RBM	HMM	RBM*	IT-RBM	HMM	RBM*	IT-RBM
accuracy	0.6723	0.9702	**0.9734**	0.8046	0.8629	**0.9296**	0.7222	0.8547	**0.8547**
F1-score	0.7719	0.9708	**0.9740**	0.8768	0.8915	**0.9470**	0.7619	0.7952	**0.8632**

* RBM is the proposed multi-value RBM

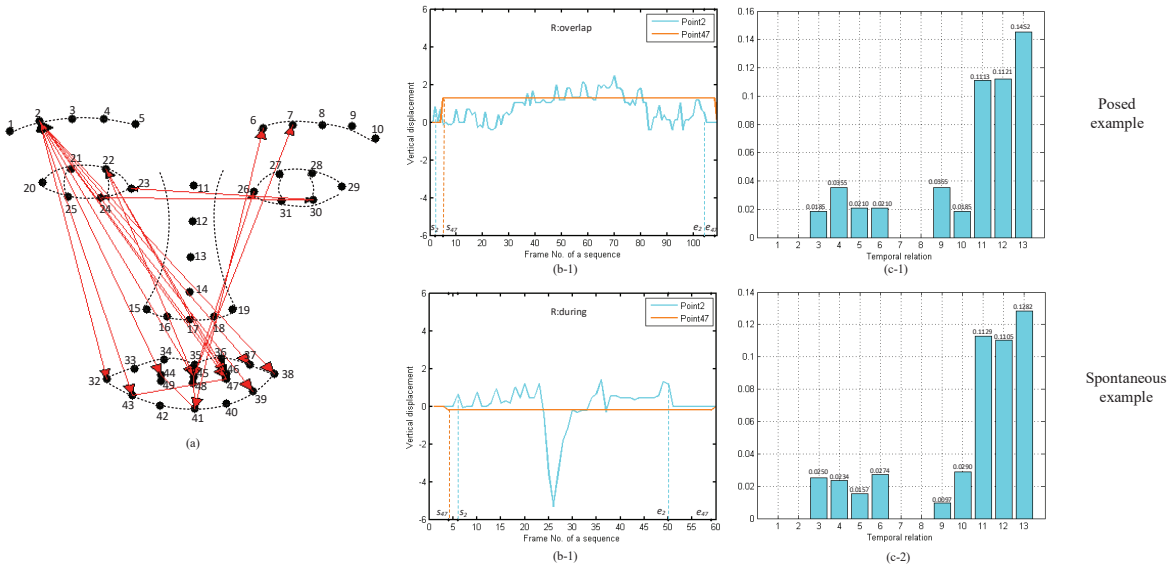

Figure 4: (a) Graphical depiction of temporal relations selected in UvA-NEMO. (b) an example of relation between point 2 and 47. (c) frequencies of thirteen relations between point 2 and point 47 with respect to posed and genuine expressions. X-axis represents the index of relationships.

13 complex temporal relations through incorporating interval algebra and can model global temporal relations. These result in better performance for posed and spontaneous expression distinction.

To further demonstrate the effectiveness of the proposed IT-RBM in capturing spatial and temporal patterns inherent in posed and spontaneous expressions, in Fig. 4 we graphicly depict the selected primitive event pairs and temporal relations on the UvA-NEMO database. Fig. 4 (a) is the 40 selected event pairs. We can find most links involved with points around eyebrow, eyepit, and lips. Ekman et al.'s work [7] [8] indicated that orbicularis oculi and the zygomatic major are most related muscles in posed and spontaneous expressions. Our findings is inconsistent with their observations.

Fig. 4 (b-1) and (b-2) show an example of temporal relations between point 2 and point 47 in posed and spontaneous expression, and the histogram (c-1) and (c-2) are the frequencies of all the thirteen relations between feature point 2 and 47 in two expressions. We can find from Table 1 that in relation 3,6 and 11, $ts_2 - ts_{47} > 0$ means that event v_{47} start before event v_2, while in relation 4,5 and 12, $ts_2 - ts_{47} < 0$ means that event v_2 start before v_{47}. What's more we can find from histogram, in posed expression the frequencies of relation 4 and 12 are higher than relation 3 and 11, and in spontaneous expression the frequencies of relation 3, 6 and 11 are higher than relation 4, 5, and 12, these indicate that in posed expression v_2 start before v_{47} in more cases, in genuine expression v_{47} start before v_2 in more cases. Furthermore, point 2 and 47 represent the muscle of right eyebrow and left lip corner singly, we can draw a conclusion that posed expression begin on the right face and spontaneous expression begin on the left face in most cases. This is in accordance with Ross and Pulusu's [16] research.

3.3 Comparison with related work

In our work, we compare the proposed method with two kinds of methods, i.e, model-based and feature-driven methods. Current model-driven methods of posed and spontaneous expression distinction mainly conducted experiments on the NVIE database and the SPOS database. Since the NIVE database only provide the onset and the apex frames of the posed expressions, we can not conduct experiments using IT-RBM on the NIVE database. We compared the performance of the proposed IT-RBM with current model-driven methods on the SPOS database as shown in Table. 4.

Table 4: Comparison with related work of posed and spontaneous expressions distinction on UvA-NEMO and SPOS

Method	UvA-NEMO	SPOS
Cohn vs. Schmidt[2]	0.7726	0.7250
Dibeklioglu *et al.*[3]	0.8702	0.7500
Dibeklioglu *et al.*[4]	0.9290	0.7875
Wu *et al.*[24]	0.9140	0.7950
Wu *et al.*[25]	0.9395	0.8125
Wang *et al* [21]	-	0.7479
Wang *et al.*[22]	-	0.7607
Quan *et al.*[9]	-	0.7607
IT-RBM	**0.9734**	**0.8632**

From Table.4, we find that the proposed IT-RBM outperforms state-of-the-art model based methods. Since Wang *et al.* [22] proposed to use RBM to model spatial patterns for posed and spontaneous expression distinction. Quan *et al.* [9] proposed latent regression Bayesian network (LRBN) to capture the spatial patterns for posed and genuine expressions distinguishing. The current model based methods only capture spatial patterns, while the proposed IT-RBM jointly model spatial and temporal patters. Therefore, the proposed IT-RBM can fully represent posed and spontaneous expressions, and lead to superior performance.

On the UvA-NEMO smile database, we compare the proposed method with state-of-the-art feature-driven methods. The DISFA+ dataset is a new database, which is available for research purpose just from June 2016. Till now, no work performs posed and spontaneous expression distinction on this database. Therefore, we can not compare with other work on this database. The experimental results are shown in the Table 4. We find on the UvA-NEMO database, the accuracy of IT-RBM is higher than result of Wu *et al.* [25] which is the best performance of current state-of-the-art feature-driven methods. They adopted completed LBP features from three orthogonal planes and extracted robust and discriminative patterns to classify posed and spontaneous expressions. Although current feature-driven methods model the inherent spatial and temporal patterns to some extent through defining discriminative features, they can not fully exploiting spatial and temporal patterns embedded in posed or spontaneous expressions through defined features. But the proposed IT-RBM can successfully represent the spatial and temporal patterns through its structure and parameters, and thus achieves best performance.

4 CONCLUSION

In this paper, we proposed a novel dynamic temporal relation and spatial structure based method termed as interval temporal restricted Boltzmann machine(IT-RBM), jointly capture global spatial patterns and complex temporal patterns embedded in posed expressions and spontaneous expressions respectively for distinguishing two expressions. We consider a facial expression as a complex activity that consists of temporally overlapping or sequential primitive facial events, which are defined as the motion of feature points. Moreover, we introduce Allen's Interval Algebra to depict more complex temporal relations between primitive events through a two-layer Bayesian network of bottom part of IT-RBM. Furthermore, we propose a multi-value RBM to capture intrinsic global spatial patterns among facial events. An IT-RBM contains three layers, hidden nodes layer the visible facial primitive event nodes layer and the temporal relation nodes layer. During training, we propose efficient learning algorithm to learn spatial patterns and temporal patterns simultaneously through maximum log likelihood. During testing, the samples are classified into posed or spontaneous expressions according to the IT-RBM with the larger likelihood. We extended annealing importance sampling to IT-RBM for calculating partition function of multi-value RBM.

To evaluate the performance, we conduct the experiment on three benchmark datasets, the results demonstrate the ability of the proposed method in exploiting intrinsic global spatial-temporal patterns in facial behavior as well as its advantage over existing methodologies for posed and spontaneous expression distinction. Moreover, compared to similar method ITBN for expression recognition, the performance indicated that IT-RBM is not only appropriate for spontaneous and posed expressions classification, but also can be applied to different facial expression recognition.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.This work has been supported by the National Science Foundation of China (Grant No. 61473270, 61228304, 61175037), and the project from Anhui Science and Technology Agency (1508085SMF223).

REFERENCES

[1] James F Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 11 (1983), 832–843.
[2] Jeffrey F Cohn and Karen L Schmidt. 2004. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing* 2, 02 (2004), 121–132.
[3] Hamdi Dibeklioğlu, Albert Salah, and Theo Gevers. 2012. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. *Computer Vision–ECCV 2012* (2012), 525–538.
[4] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2015. Recognition of genuine smiles. *IEEE Transactions on Multimedia* 17, 3 (2015), 279–294.
[5] Hamdi Dibeklioglu, Roberto Valenti, Albert Ali Salah, and Theo Gevers. 2010. Eyes do not lie: Spontaneous versus posed smiles. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 703–706.
[6] Paul Ekman. 2003. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences* 1000, 1 (2003), 205–221.
[7] Paul Ekman and Wallace V Friesen. 1982. Felt, false, and miserable smiles. *Journal of nonverbal behavior* 6, 4 (1982), 238–252.
[8] Paul Ekman, Joseph C Hager, and Wallace V Friesen. 1981. The symmetry of emotional and deliberate facial actions. *Psychophysiology* 18, 2 (1981), 101–106.
[9] Quan Gan, Siqi Nie, Shangfei Wang, and Qiang Ji. 2017. Differentiating Between Posed and Spontaneous Expressions with Latent Regression Bayesian Network. In *Proceedings of the*

Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 4039–4045. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14673

[10] Geoffrey Hinton. 2010. A practical guide to training restricted Boltzmann machines. *Momentum* 9, 1 (2010), 926.

[11] Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation* 14, 8 (2002), 1771–1800.

[12] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. 2009. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing* 27, 12 (2009), 1797–1803.

[13] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. 2016. Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–8.

[14] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. 2011. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 868–875.

[15] William H Press, Brian P Flannery, Saul A Teukolsky, William T Vetterling, and Peter B Kramer. 1987. Numerical recipes: the art of scientific computing. (1987).

[16] Elliott D Ross and Vinay K Pulusu. 2013. Posed versus spontaneous facial expressions are modulated by opposite cerebral hemispheres. *Cortex* 49, 5 (2013), 1280–1291.

[17] Ruslan Salakhutdinov and Iain Murray. 2008. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*. ACM, 872–879.

[18] Karen L Schmidt, Sharika Bhattacharya, and Rachel Denlinger. 2009. Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Journal of nonverbal behavior* 33, 1 (2009), 35–45.

[19] Melinda Seckington. 2011. Using dynamic bayesian networks for posed versus spontaneous facial expression recognition. *Mater Thesis, Department of Computer Science, Delft University of Technology* (2011).

[20] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. 2006. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 162–170.

[21] Shangfei Wang, Chongliang Wu, Menghua He, Jun Wang, and Qiang Ji. 2015. Posed and spontaneous expression recognition through modeling their spatial patterns. *Machine Vision and Applications* 26, 2-3 (2015), 219–231.

[22] Shangfei Wang, Chongliang Wu, and Qiang Ji. 2016. Capturing global spatial patterns for distinguishing posed and spontaneous expressions. *Computer Vision and Image Understanding* 147 (2016), 69–76.

[23] Ziheng Wang, Shangfei Wang, and Qiang Ji. 2013. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3422–3429.

[24] Pingping Wu, Hong Liu, and Xuewu Zhang. 2014. Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 1240–1244.

[25] Pingping Wu, Hong Liu, Xuewu Zhang, and Yuan Gao. 2016. Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics. *Frontiers* 1 (2016).

[26] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.