A Novel Dynamic Model Capturing Spatial and Temporal Patterns for Facial Expression Analysis

Shangfei Wang, *Senior Member, IEEE*, Zhuangqiang Zheng, Shi Yin, Jiajia Yang, and Qiang Ji, *Fellow, IEEE*

Abstract— Facial expression analysis could be greatly improved by incorporating spatial and temporal patterns present in facial behavior, but the patterns have not yet been utilized to their full advantage. We remedy this via a novel dynamic model-an interval temporal restricted Boltzmann machine (IT-RBM) - that is able to capture both universal spatial patterns and complicated temporal patterns in facial behavior for facial expression analysis. We regard a facial expression as a multifarious activity composed of sequential or overlapping primitive facial events. Allen's interval algebra is implemented to portray these complicated temporal patterns via a two-layer Bayesian network. The nodes in the upper-most layer are representative of the primitive facial events, and the nodes in the lower layer depict the temporal relationships between those events. Our model also captures inherent universal spatial patterns via a multi-value restricted Boltzmann machine in which the visible nodes are facial events, and the connections between hidden and visible nodes model intrinsic spatial patterns. Efficient learning and inference algorithms are proposed. Experiments on posed and spontaneous expression distinction and expression recognition demonstrate that our proposed IT-RBM achieves superior performance compared to state-of-the art research due to its ability to incorporate these facial behavior patterns.

Index Terms—Interval temporal restricted Boltzmann machine, global spatial and temporal patterns, posed and spontaneous expressions distinction, expressions categories recognition

1 Introduction

There has been a proliferation of research on facial expression analysis recently, since facial expression is a crucial channel for both human-human communication and human-robot interaction.

Current works on facial expression recognition may be categorized into either of two approaches: a frame-based approach or a sequence-based approach. A frame-based approach recognizes facial expressions from static facial images, usually from the manually annotated apex frame. This approach completely disregards the important dynamic patterns inherent in facial behavior. A sequence-based approach relies on the whole image sequence, and thus has the potential to model both spatial and temporal patterns through features or dynamic classifiers. Current works either employ hand-crafted spatial and temporal features or use learned representation through deep networks. Several dynamic classifier models, such as hidden Markov models (HMMs), dynamic Bayesian networks (DBNs), latent conditional random fields (LCRFs), long short-term memory networks (LSTMs), or gated recurrent unit networks (GRUs), are frequently used. All of these works try to find more discriminative features or more powerful classifiers to explore embedded spatial and temporal patterns, and have been successful for facial expression analysis. We refer to these approaches as feature-driven methods.

Few works consider the underlying anatomic mechanisms governing facial muscular interactions. Nearly any facial expression can be deconstructed into the contraction or relaxation of one or more facial muscles. These facial muscle movements interact in space and time to convey different expressions. At each time slice, facial muscle motions may co-occur or be mutually exclusive. For example, as shown in Figs. 1a and 1b, most people raise the inner brow and outer brow simultaneously, since both motions are related to the frontalis muscle group. The lip corner puller rarely occurs in tandem with the lip corner depressor, as shown in Figs. 1c and 1d. The lip corner puller uses the muscle group zygomaticus major, and the latter is produced by the depressor anguli oris muscle group. Temporally, the movement of one facial muscle can either activate, meet, overlap, or succeed another muscle. As shown in Fig. 2, for example, most people show happiness by stretching their mouths while raising their cheeks. Therefore, the contraction of zygomatic major is more likely to occur asymmetrically if a smile is posed rather than spontaneous. When an expression is natural and spontaneous, the trajectory is typically smoother. It has a shorter duration, and onset is gradual rather than immediate. Such spatial and temporal patterns caused by the interaction of facial expression muscles are extremely complex, time-dependent, and global, yet have not been fully modeled by current facial expression analysis methods.

- *S. Wang, Z. Zheng, S. Yin, and J. Yang are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China. E-mail: sfwang@ustc.edu.cn, {zheng001, davidyin, yang25}@mail.ustc.edu.cn.*
- *Q. Ji is with the Department of Electrical, Computer, Systems Engineering Rensselaer Polytechnic Institute, Troy, NY 12180-3590. E-mail: qji@ecse.rpi.edu.*

Manuscript received 3 Oct. 2017; revised 9 Mar. 2019; accepted 13 Apr. 2019. Date of publication 17 Apr. 2019; date of current version 4 Aug. 2020. (Corresponding author: Jiajia Yang.) Recommended for acceptance by E. G. Learned-Miller. Digital Object Identifier no. 10.1109/TPAMI.2019.2911937

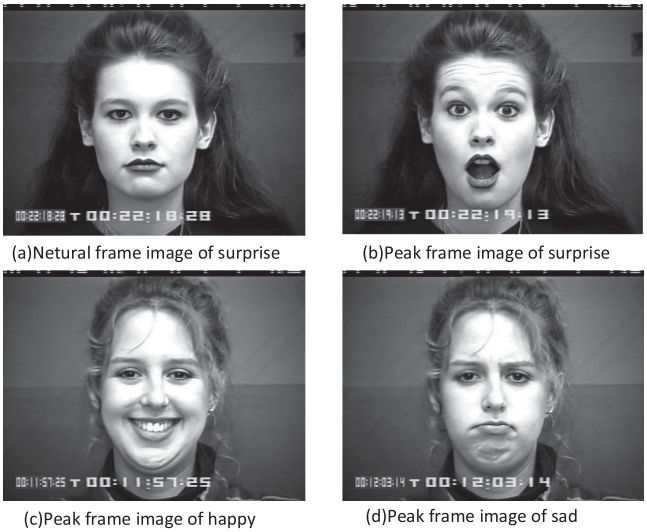


Fig. 1. Sample images demonstrating spatial patterns inherent in expressions.

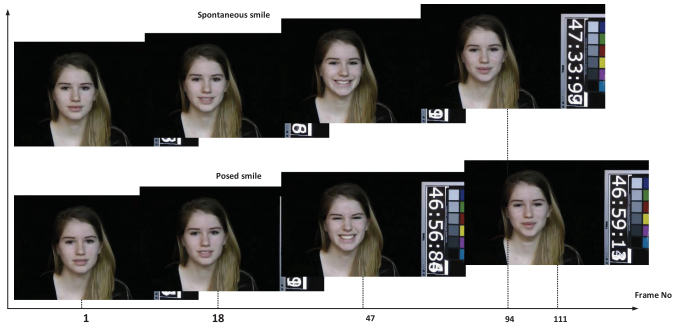


Fig. 2. Image sequences demonstrating temporal patterns inherent in expressions. The x-axis is the frame number.

We propose a novel dynamic model that leverages the complex spatial and temporal patterns caused by the underlying anatomic mechanism for expression analysis. We assume an expression is a multifarious activity made up of sequential or overlapping primitive facial events, and that each event takes place over a certain amount of time. First, we introduce Allen's interval algebra to capture several types of temporal relationships, including A takes place before B, A meets B, A overlaps with B, A initiates B, A occurs during B, A finishes B, A is equal to B, and the inverse of the first six relations. We implement the complex temporal relations using a Bayesian network incorporating primitive facial event nodes and temporal relationship nodes. The links connecting the two types of nodes characterize their temporal relationships. Next, a restricted Boltzmann machine (RBM) is adopted to represent the global spatial patterns among primitive facial events. The visible nodes of the RBM depict primitive facial events, and the connections between hidden nodes and visible nodes model the spatial patterns inherent in expressions. During training, we build an IT-RBM model for each type of expression, and the parameters and structures of the proposed IT-RBM are learned through maximum likelihood. When testing occurs, the test sample label is equivalent to the model with the largest log likelihood.

The proposed IT-RBM differs from other dynamic models in that it introduces Allen's interval algebra to capture all 13 temporal relations. Unlike current dynamic models, which are limited to time-slice structure and must assume stationary and time-independent temporal relations, the suggested model can capture more complex global temporal relationships.

The paper is organized as follows. Section 2 is a brief review of related works on expression analysis, including expression recognition as well as posed and spontaneous expression distinction. Section 3 details the proposed IT-RBM model. Section 4 outlines the experiments and analysis, with posed and spontaneous expression distinction experiments outlined in Section 4.1 and expression recognition experiments detailed in Section 4.2. Section 5 summarizes our work.

2 Related Work

2.1 Posed and Spontaneous Expression Distinction

Inner feelings may be disguised with a posed expression, but true emotions are conveyed via spontaneous expressions. It is difficult to distinguish one from the other, since expressions vary by subject and condition and the differences between a spontaneous and posed expression are subtle. The inherent spatial and temporal patterns in facial expressions can be leveraged to improve distinction between these similar types of expressions.

Behavioral research has found slight but distinctive differences between temporal and spatial patterns in posed and spontaneous expressions. Examples of temporal patterns include speed, trajectory, amplitude, and duration of expression onset and offset. For example, Ekman et al. [1], [2] found that spontaneous expressions usually have a smoother trajectory and shorter duration than posed expressions. Schmidt et al. [3] revealed that for posed smiles, the maximum speed of movement onset is greater than it is for spontaneous smiles. Deliberate eyebrow raises are shorter in duration and have a greater maximum speed and amplitude than spontaneous, natural eyebrow raises. Spatial patterns mainly consist of the movement of facial muscles. For example, Ekman et al.' s work [1] found that the orbicularis oculi only contract during spontaneous smiles. When smiling, the contraction of the zygomatic major muscle is more likely to be asymmetric for a posed expression than a spontaneous one [4]. Ross and Pulusu [5] indicated that posed expressions typically commence on the right side of the face, while spontaneous expressions originate on the left side of the face. This is especially true for upper facial expressions. Namba et al.'s work [6] compared the morphological and dynamic properties of spontaneous and posed facial expressions as they related to surprise, amusement, disgust, and sadness. For amusement, AUs yield no significant differences. For disgust, AU10 and AU12 occur more frequently when an expression is spontaneous rather than posed, while AU17 appears more often in posed expressions. For sadness, morphological properties of spontaneous facial expressions are not observed, while AU4, AU7, and AU17 are most frequently observed in posed facial expressions.

Most research uses certain features to distinguish between posed and spontaneous expressions. Cohn and Schmidt [7] adopted temporal features, including duration, amplitude, and the ratio between the two. Valstar [8] utilized features such as speed, duration, trajectory, intensity, symmetry, and

the occurrence order of brow actions based on fiducial facial point displacement. Dibeklioglu et al. [9] described the dynamics of eyelid, cheek, and lip corner movements using amplitude, duration, speed, and acceleration. Seckington [10] represented temporal dynamics using six features (i.e., morphology, apex overlap, symmetry, total duration, onset speed, and offset speed).

Static classifiers (e.g., linear discriminant classifiers [7], support vector machines [11], k-NN [12], and naive Bayesian classifiers [12]) and dynamic classifiers (e.g., continuous hidden Markov models [12] and dynamic Bayesian networks [10]) were investigated for the task of distinguishing between posed and spontaneous expressions. Static classifiers model the mapping between features and expression types, while dynamic classifiers model the temporal relationships.

Progress has been made in distinguishing between posed and spontaneous expressions. However, these feature-driven methods do not explicitly leverage the underlying interactions between facial expression muscles, and their influences on posed and spontaneous expressions.

Recently, Wang et al. [13] proposed a model-based method using multiple Bayesian networks (BNs) to capture spatial patterns for expressions given gender and expression categories. This model only includes local dependencies due to the first-order Markov assumption of BNs; it cannot capture high-order or global relations. Wu et al. [14] proposed to address that issue by implementing a restricted Boltzmann machine to explicitly model complex joint distributions over feature points. RBMs introduce a layer of latent units, allowing them to model high-order dependencies among variables [15]. Although this model is an improvement, it does not leverage the dependencies among hidden units. Quan et al. [16] employed a latent regression Bayesian network (LRBN) to leverage higher-order and global dependencies among facial features. A latent regression Bayesian network differs from an RBM in that it is a directed rather than undirected model. The "explaining away" effect in Bayesian networks allows LRBNs to capture dependencies among both latent and visible variables; these dependencies are vital to accurately represent the data. The success of each of these three model-based works prove that spatial patterns can contribute to the differentiation of posed from spontaneous expressions.

Thus far, there have not been many attempts to capture and leverage the spatial and temporal patterns embedded in posed and spontaneous facial expressions. We propose an interval temporal restricted Boltzmann machine (IT-RBM) to jointly capture global and complex patterns and improve the task of expression distinction.

2.2 Expression Recognition

Expression recognition has attracted much more attention than the distinction between posed and spontaneous expressions. Corneanu et al.'s work [17] and Brais Martinez et al.'s work [18] provided a literature review of facial expression recognition.

Mainstream facial expression recognition works regard facial expression recognition as a pattern recognition problem and focus primarily on discriminative features and powerful classifiers. For features, both engineered dynamic representations and learned representations from video volumes are exploited to encode temporal variations among sequence frames. Engineered dynamic representations such as LBP-TOP [19], and Gabor motion energy [20] do not require labelled sequences for training, and thus are simple and generic for any expression analysis tasks. However, the optimality is questionable. The learned representations may attain state of the art performance, but they require many training videos with ground-truth labels. Dynamic graphic models, such as HMM [21], [22], [23] and DBN [24], have commonly been used for facial expression analysis tasks. As time-slice (based on time points) graphical models, these dynamic models represent each activity as a sequence of events occurring instantaneously, and thus offer three time-point relations (i.e., before, follows and equals). Since facial expressions are complex and consist of facial events that may be sequential or temporally overlapping, current dynamic graphical models are unable to represent several of the temporal relations happening between events throughout the activity. Recently, deep dynamic models such as LSTM [25] have been adopted for facial expression analysis. Usually, a convolutional neural network (CNN) is used to obtain static representations from each frame, and then the learned static representations are fed into the LSTM to learn dynamic representations and expression classifiers simultaneously. In spite of its good performance, LSTM requires a lot of training data. Furthermore, LSTM is also a time-slice model and cannot successfully represent the global and complex temporal relations between primitive facial events inherent in facial expressions. Just as in expression distinction, these feature-drive expression recognition methods ignore the underlying interaction among facial expression muscles.

A facial expression is defined as at least one motion of the facial muscles over a period of time. These muscle movements commonly appear in certain patterns to communicate different expressions. For example, the facial expression of happiness is characterized by raised cheeks and a stretched mouth. Surprise is usually displayed by widened eyes and a gaping mouth. A look of sadness is easily identified by upwardly slanted eyebrows and a frown. An expression of anger is often determined by squeezed eyebrows as well as tight and straight eyelids. Fear typically includes widened eyes and eyebrows slanted upward. These expression-dependent temporal and spatial patterns are essential for expression recognition, but have yet to be exploited thoroughly.

As far as we know, only one related work attempts to capture these patterns for expression recognition. Wang et al. [26] suggested an interval temporal Bayesian network (ITBN) including temporal entity nodes and temporal relation nodes. The links between the former types of nodes represent spatial dependencies among temporal entities. Links joining temporal relation nodes with their corresponding temporal entities are representative of the temporal relationships between the two connected temporal entities. Thus, the ITBN is able to leverage spatial and temporal patterns. Due to the Markov assumption, Bayesian networks only capture local dependences. Therefore, instead of a BN, we employ an RBM to capture and depict global spatial patterns. Since Allen's interval algebra defines complete temporal relations between two events and a BN can fully capture dependencies between two events, we still employ a BN to model temporal patterns embedded in expression changes.

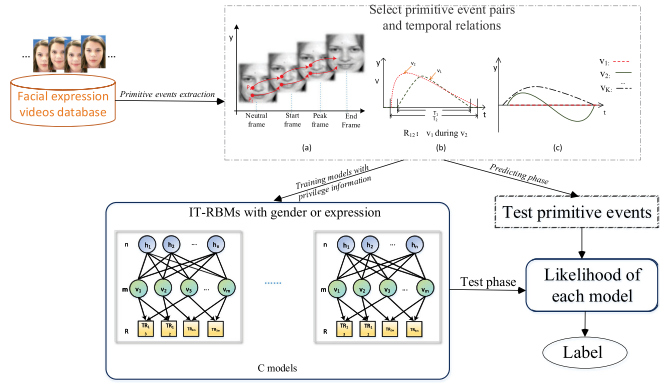


Fig. 3. Outline of the recognition system.

The proposed IT-RBM is a novel dynamic model and can provide complex and global relations through the use of interval algebra, which defines complete temporal relations between two events. This is an improvement over typical dynamic models like HMM and DBN, which use a time-slice structure to present three time-point relations, and are only able to capture stationary dynamics.

This paper makes the following contributions to this field of study:

1. A novel dynamic model, IT-RBM, is proposed to jointly capture both global spatial patterns and complex temporal patterns.
2. We explicitly model spatial-temporal patterns innate to various expression categories for expression recognition.
2. We explicitly model spatial-temporal patterns found in posed and spontaneous expressions to better distinguish between those expressions.

A previous version of the paper appeared as Yang et al.' work [27], which proposed an IT-RBM to capture and utilize spatial and temporal patterns embedded in posed and spontaneous expressions for expression distinction. Unlike the previous version, which only focuses on posed and spontaneous expression distinction, this paper extends the proposed IT-RBM for expression recognition. To show the effectiveness of the proposed IT-RBM for expression recognition, experiments are conducted on the CK+ and the MMI databases. We have added two models for posed and spontaneous expression distinction (i.e., PS_gender model and PS_exp model), since the spatial and temporal patterns embedded in expressions are influenced by gender and type of expression. For the PS_gender model, we train four models from male posed, male spontaneous, female posed, and female spontaneous samples. For the PS_exp model, we train a posed model and a spontaneous model for each expression type.

3 Proposed Method

Facial expressions are the results of a set of muscle movements over a period of time. At each time slice, facial muscle motions can co-occur or be mutually exclusive. From a temporal perspective, the movement of one facial muscle can activate, overlap, or follow the movement of another muscle. Because of the difficulty in measuring minute facial muscle motions, the movements of facial feature points are used to define primitive facial events as recommended by Wang et al.' work [26]. Each feature point movement is a singular primitive facial event. The interval relation between each pair of events can be defined as one of 13 interval relations proposed by Allen's interval algebra [28]. First, we select the primitive event pairs with the largest interval relation variance among the different expressions. For each type of expression, an IT-RBM model is constructed using the selected events and interval relations. The global spatial and temporal patterns are jointly captured during training. During testing, the label of a test sample is the model with the largest likelihood. The method framework is illustrated in Fig. 3.

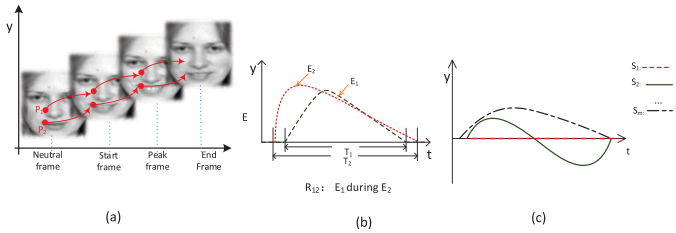


Fig. 4. (a) Facial muscle movement as captured by the movement of facial points. (b) Duration for events V_1 and V_2 and their temporal relations. (c) Example movement states of a primitive facial event.

This section focuses first on the extraction of primitive facial events. Then, we describe the definition and selection of temporal relations. After that, the proposed IT-RBM model is presented in detail.

3.1 Primitive Facial Events Extraction

Given the data set of sample videos with different types of expression, denoted as $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$, where $x^{(i)}$ is the ith video with frame length of f_i, and $y^{(i)} \in \{0, 1, \ldots, C\}$ is its expression label (C is the total expression classes of videos), N is the number of total sample videos. Each frame is a facial image with P_{no} facial points. Each video is assumed to contain primitive facial events, which are either sequential or temporally overlapping. A primitive facial event is the movement of one feature point and includes the motion state, the commencement time when the feature point is no longer in neutral position, and the moment when the point returns to neutral. Fig. 4 depicts a primitive event corresponding to ith facial point denoted as $V_i = \langle ts_i, te_i, v_i \rangle (ts_i, te_i \in \mathbb{R}, ts_i < te_i, v_i \in \{1, 2, \ldots, K\})$, ts_i and te_i are representative of the start and end times respectively, v_i represents the motion state, $\{1, 2, \ldots, K\}$ represents all possible primitive event states. As expression videos differ in frame length, we normalize all frames to the shortest frame length in the training set len. Samples are equidistantly down-sampled to len. We obtain K movement states by using K-means clustering on the feature point displacement sequence.

Fig. 4 illustrates some example primitive facial events. In (a) facial points P_1 and P_2 correspond to events V_1 and V_2, representing the muscle motions of the right wing of the nose and the right mouth corner respectively. (b) shows the trace of V_1 and V_2 along the vertical direction, and T_1, T_2 are their corresponding durations. Each event has K possible states representing its movement pattern throughout the duration, as (c) shows. The flat red line depicts a point that remains in neutral for the entire process, and the other states

TABLE 1
TR 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	$= 0$	$= 0$	–	–	

The horizontal bars represent the time interval of the corresponding primitive events.

are representative of $k-1$ movement patterns. For example, a point that moves up and then returns to neutral would be represented by state S_m(shown as the dotted black line). A more complex pattern is depicted by state S_2 (the solid line), in which a point moves upward and then downward.

3.2 Temporal Relations Definition and Selection

According to Allen's interval algebra [28], there are 13 potential temporal relationships between two primitive events as illustrated in Table 1. The 13 possible relations $\mathbb{I} = \{b, bi, m, mi, o, oi, s, si, d, di, f, fi, eq\}$, representing before, meets, overlaps, starts, during, finishes, equals and their inverses. The temporal relationships between pairs of facial events V_i and V_j can be obtained by calculating the temporal distance $dis(V_i, V_j)$ according to Eq. (1).

$$dis(V_i, V_j) = [ts_i - ts_j, ts_i - te_j, te_i - ts_j, te_i - te_j]. \tag{1}$$

The extraction of primitive facial events yields $P_{no} * (P_{no}-1)$ pairs of events and the corresponding temporal relations for each sample. It is expected that discriminative temporal relations will have a wider variance between expression types, so we propose a Kullback-Leibler divergence-based score [29] to measure the difference between the two probability distributions. The score of event pair V_i, V_j is defined in Eq. (2), where TR_{ij} represents relation between primitive event pair V_i, V_j, $P_x(TR_{ij})$ and $P_y(TR_{ij})$ are the probability distribution of TR_{ij} for the x and y expression respectively. D_{KL} stands for the KL divergence. Primitive event pairs are ranked by the score, and the top ξ pairs with m events are selected.

$$S_{ij} = \sum_{x,y \in \{1,2,\ldots,C\}} \left(D_{KL}(P_x(TR_{ij}) \| P_y(TR_{ij})) + D_{KL}(P_y(TR_{ij}) \| P_x(TR_{ij}))\right). \tag{2}$$

3.3 Capturing Spatial and Temporal Patterns Through the IT-RBM Model

Our proposed hybrid graphic model known as IT-RBM is shown in Fig. 5. The upper section is a multi-value RBM and the lower layer is a Bayesian network. The uppermost layer contains n binary latent variables $h_j \in \{0, 1\} (j \in \{1, 2, \ldots, n\})$. The layer below that contains m visible nodes. $v_i \in \{1, \ldots, K\} (i \in \{1, 2, \ldots, m\})$ describes m selected facial events. Each facial event consists of K motion states represented by an one-hot vector. Specifically, v_i consists of binary nodes $v_{i1}, v_{i2}, \ldots, v_{iK}$, thus $v_i = k$ can be represented with an one-hot vector by setting $v_{ik} = 1$, other K-1 binary nodes to zeros. The bottom layer contains ξ temporal relation nodes, $TR \in \mathbb{I}$ representing 13 temporal relations. Complex temporal relations are captured by the lower part; the spatial dependencies among facial events are modeled by the upper part. Eq. (3) shows the joint probability of the suggested model.

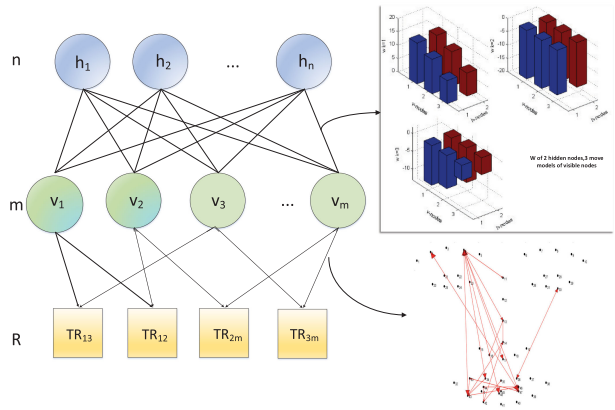


Fig. 5. An example of IT-RBM model.

$$P(v, TR) = P(TR|v)P(v) = P(TR|v) \sum_h P(v, h), \tag{3}$$

where

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

TR_r represents the rth temporal relation node. $\pi(TR_r)$ are the two primitive event nodes that produce TR_r.

After primitive events and temporal relations are extracted, given training data $D_t = \{(v^{(1)}, TR^{(1)}), (v^{(2)}, TR^{(2)}), \ldots, (v^{(N_t)}, TR^{N_t})\}$, where N_t indicates the number of training samples for one expression, $v^{(i)}$ and $TR^{(i)}$ represents motion states and temporal relations of ith sample. The goal of model learning is log likelihood maximization, shown as follows:

$$\theta^* = \arg\max_{\theta} \frac{1}{N_t} \sum \left(\log P(v; \theta) + \log P(TR|v; \theta)\right). \tag{5}$$

Eq. (5) demonstrates that we can factorize the log likelihood of IT-RBM into the sum of the log likelihood of RBM and the log likelihood of BN. Since the model parameters of RBM θ_{RBM} is independent of model parameters of BN θ_{BN}, we can train RBM and BN separately. Training of the multi-value RBM only concerns motion states of primitive event, so we denote $D_t^{RBM} = \{v^{(1)}, v^{(2)}, \ldots, v^{(N_t)}\}$.

The marginal distribution of the visible units is calculated as Eq. (6),

$$\begin{aligned} P(v) &= \sum_h P(v, h) = \frac{1}{Z} \sum_h e^{-E(v,h)} \\ &= \frac{1}{Z} e^{\sum_{i=1}^{m} \sum_{k=1}^{K} b_{ik} v_{ik}} \prod_{j=1}^{n} (1 + e^{a_j + \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik}^{j} v_{ik}}), \end{aligned} \tag{6}$$

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

$$E(v,h) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} v_{ik} w_{ik}^j h_j - \sum_{i=1}^{m}\sum_{k=1}^{K} v_{ik} b_{ik} - \sum_{j=1}^{n} h_j a_j. \tag{7}$$

The gradient with respect to $\theta_{RBM} = \{W, a, b\}$ can be calculated as Eq. (8), where $P(v,h)$ and $P(h|v)$ denote the model-defined distribution, v in the first term is from training set, v in the second term is sampled from model-defined distribution $P(v)$.

$$\begin{aligned}\Delta\theta_{RBM} &= \varepsilon \frac{\partial \log P(v)}{\partial \theta_{RBM}} \\ &= \varepsilon\left(-\sum_h P(h|v)\frac{\partial E(v,h)}{\partial \theta_{RBM}} + \sum_{v,h} P(v,h)\frac{\partial E(v,h)}{\partial \theta_{RBM}}\right).\end{aligned} \tag{8}$$

The contrastive divergence (CD) algorithm is used to overcome the challenge of inferring the second term of Eq. (8), which is intractable and is needed for gradient calculation [30]. The conditional distribution of visible nodes given hidden nodes and the conditional distribution of hidden nodes given visible nodes are softmax function and logistic function respectively, as follows:

$$P(v_i = k|h) = \frac{\exp(b_{ik} + \sum_{j=1}^{n} h_j w_{ik}^j)}{\sum_{l=1}^{K}\exp(b_{il} + \sum_{j=1}^{n} h_j w_{il}^j)} \tag{9}$$

$$P(h_j = 1|v) = \sigma(a_j + \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik}^j v_{ik}). \tag{10}$$

The detailed algorithm for learning multi-value RBM is shown as Algorithm 1.

Algorithm 1. The Training Algorithm for Multi-Value RBM using CD Learning

Require: Training data:$D_t^{RBM} = \{v^{(1)}, v^{(2)}, \ldots, v^{(N_t)}\}$, latent nodes number n, learning rate ε, maximum training times T
Ensure: w_{ik}^j, a_j, b_{ik}
1: Initialize: set w, a, b to small random values
2: **for** $t = 1, 2, \ldots, T$ **do**
3: sample one example v from D_t^{RBM}
4: **for** $j = 1, 2, \ldots, n$ **do**
5: Sample $h_j \sim p(h_j|v)$ with Eq. (10)
6: **end for**
7: **for** $i = 1, 2, \ldots, m$ **do**
8: Sample $v_i' \sim p(v_i|h)$ with Eq. (9)
9: **end for**
10: parameters update:
11: $w_{ik}^j \leftarrow w_{ik}^j + \varepsilon(P(h_j = 1|v)v_{ik} - P(h_j = 1|v')v_{ik}')$
12: $b_{ik} \leftarrow b_{ik} + \varepsilon(v_{ik} - v_{ik}')$
13: $a_j \leftarrow a_j + \varepsilon(P(h_j = 1|v) - P(h_j = 1|v'))$
14: **end for**

The conditional probability distributions for each temporal relation node TR_{ij} given its parent nodes v_i and v_j are used to define parameters for the BN. The structure of the BN and the number of parameters can be determined after temporal relations are selected. The goal of parameter approximation is to find the maximum likelihood estimate of parameter θ_{BN} given training data $D_t = \{(v^{(1)}, TR^{(1)}), (v^{(2)}, TR^{(2)}), \ldots, (v^{(N_t)}, TR^{N_t})\}$. This is depicted in Eq. (11).

$$\theta_{BN} = \arg\max_{\theta_{BN}} \sum_{N_t} \log P(TR|v; \theta_{BN}). \tag{11}$$

3.4 Expression Analyses

An IT-RBM model can be obtained for each expression after training. During testing, test sample t is labeled with the class that has the largest log likelihood value, according to Eq. (12). In that equation, y^* represents the predicted label and C is the number of expression categories (as well as the number of IT-RBM models).

$$y^* = \max_{y\in\{1,\ldots,C\}} \{\log P(t|\theta_y)\}. \tag{12}$$

The log likelihood that IT-RBM trained on class y assigned to test sample t is as follows:

$$\begin{aligned}&\log P(t|\theta_y) \\ &= \log\left(\sum_h \exp(-E(h,t;\theta_y))\right) - \log Z(\theta_y) + \log\left(P(TR|t;\theta_y)\right),\end{aligned} \tag{13}$$

in which the first and third terms can be directly calculated and the partition function Z is intractable. The extended AIS method inspired by annealed importance sampling (AIS) [31] is used to compute the partition function of multi-value RBM.

AIS approximates the ratio of the partition functions of the object RBM to the base-rate RBM. For example, suppose there are two multi-value RBMs with parameters $\theta_A = \{W^A, b^A, a^A\}$ and $\theta_B = \{W^B, b^B, a^B\}$. These RBMs define probability distributions P_A and P_B over the same $v \in \{0, 1, \ldots, K\}^m$, and $h^A \in \{0,1\}^{n_A}, h^B \in \{0,1\}^{n_B}$.

First, the intermediate distribution sequence for $\tau = 0, \ldots, n$ is defined as

$$P_\tau(v) = \frac{P_\tau^*(v)}{Z_\tau} = \frac{1}{Z_\tau}\sum_h \exp(-E_\tau(v,h)), \tag{14}$$

where the energy function is delineated in Eq. (15), $P_0(v) = P_A$ and $P_n(v) = P_B$. Eq. (16) approximates the unnormalized probability over visible units, where $0 = \beta_0 < \beta_1 < \ldots < \beta_\tau < \ldots < \beta_n = 1$.

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

$$\begin{aligned}P_\tau^*(v) = {} & e^{(1-\beta_\tau)\sum_i\sum_k {b_{ik}}^A v_{ik}} \prod_{j=1}^{n_A}(1 + e^{(1-\beta_\tau)(\sum_i\sum_k w_{ik}^{jA} v_{ik} + a_j^A)}) \\ & * e^{\beta_\tau\sum_i\sum_k {b_{ik}}^B v_{ik}} \prod_{j=1}^{n_B}(1 + e^{\beta_\tau(\sum_i\sum_k w_{ik}^{jB} v_{ik} + a_j^B)}).\end{aligned} \tag{16}$$

Next, we establish a Markov chain transition operator $T_\tau(v'; v)$ that leaves $P_\tau(v)$ invariant. Logistic and softmax functions yield the conditional distributions as follows:

TABLE 2
Data Distribution of SPOS and DISFA+

	SPOS		DISFA+	
Expression	P	S	P	S
Anger(An)	14	13		
Disgust(Di)	14	23	163	81
Fear(Fe)	14	32	163	63
Happy(Ha)	14	66	42	18
Sadness(Sa)	14	5	122	54
Surprise(Su)	14	11	82	36
Total	84	150	572	252

$$P(h_j^A = 1|v) = \sigma((1-\beta_\tau)(a_j^A + \sum_i \sum_k w_{ik}^{jA} v_{ik})) \tag{17}$$

$$P(h_j^B = 1|v) = \sigma(\beta_\tau(a_j^B + \sum_i \sum_k w_{ik}^{jB} v_{ik})) \tag{18}$$

$$P(v_i = k|h) = \frac{\exp\left((1-\beta_\tau)\left(b_{ik}^A + \sum_{j=1}^n h_j^A w_{ik}^{jA}\right) + \beta_\tau\left(b_{ik}^B + \sum_{j=1}^n h_j^B w_{ik}^{jA}\right)\right)}{\sum_{l=1}^K \exp\left((1-\beta_\tau)\left(b_{il}^A + \sum_{j=1}^n h_j^A w_{il}^{jA}\right) + \beta_\tau\left(b_{ik}^B + \sum_{j=1}^n h_j^B w_{il}^{jA}\right)\right)}. \tag{19}$$

Hidden units h_A and h_B are stochastically activated using Eqs. (17) and (18). A new sample is drawn using Eq. (19).

Finally, with initial $\theta_A = \{0, b^A, 0\}$, Z^A is calculated as Eq. (20). Then we can calculate Z^B with Eq. (21). We define $\omega^{(i)}$ in Algorithm 2 to avoid using too many symbols.

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

$$\frac{Z_B}{Z_A} \approx= \frac{1}{M_r} \sum_{i=1}^{M_r} \omega^{(i)} = \widehat{r}_{AIS}. \tag{21}$$

The detailed algorithm is outlined below in Algorithm 2.

Algorithm 2. The AIS Algorithm for Capturing Partition Function Z [31]

Require: Base-rate RBM's parameters $\theta_A = \theta_0$, Objective RBM's parameters $\theta_B = \theta_1$,
Ensure: Objective RBM's Z_B
1: **for** i = 1 to M_r **do**
2: **for** $\beta = 0$ to 1 **do**
3: Generate $v_1, v_2, \ldots, v_\tau, \ldots, v_n$ using T_τ as follows:
4: Sample v_1 from $P_A = P_0$
5: Sample v_2 given v_1 using T_1
6: ...
7: Sample v_τ given $v_{\tau-1}$ using $T_{\tau-1}$
8: ...
9: Sample v_n given v_{n-1} using T_{n-1}
10: **end for**
11: $\omega^{(i)} = \frac{P_1^*(v_1)}{P_0^*(v_1)} \frac{P_2^*(v_2)}{P_1^*(v_2)} \cdots \frac{P_\tau^*(v_\tau)}{P_{\tau-1}^*(v_\tau)} \cdots \frac{P_n^*(v_n)}{P_{n-1}^*(v_n)}$
12: **end for**
13: $\widehat{r}_{AIS} = \frac{1}{M_r} \sum_{i=1}^{M_r} \omega^{(i)}$
14: $Z_B = Z_A * \widehat{r}_{AIS}$

4 Experiments

The proposed IT-RBM model can be applied to distinction between posed and spontaneous expression as well as expression recognition. Therefore, to validate the proposed IT-RBM model, we conduct experiments on posed and spontaneous expression distinction as well as expression recognition.

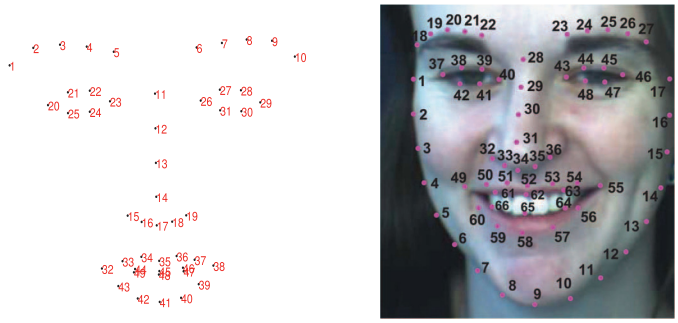


Fig. 6. Facial feature points. Left: SPOS, CK+, MMI; right: DISFA+.

4.1 Posed and Spontaneous Expression Distinction Experiments

4.1.1 Experimental Conditions

We use two benchmark databases to conduct experiments on posed and spontaneous expression distinction: the Extended DISFA (DISFA+) database [32] and the SPOS database [33]. The DISFA+ database is composed of 572 posed expression videos and 252 spontaneous expression videos. Disgust, fear, happiness, sadness, and surprise are exhibited by 9 young adults (4 male and 5 female). The SPOS database contains 84 posed expression samples and 150 spontaneous expression samples. This database explores the same expression categories as the DISFA+ database, with the addition of anger. Expressions are made by 7 subjects (4 male and 3 female). The data distribution of these databases is show in Table 2.

We extracted facial feature points from images to collect facial events as defined in Section 3.1. The supervised descent method (SDM) [34] extracts 49 facial feature points from the SPOS database, as seen in the left side of Fig. 6. The DISFA+ database provides 68 feature points extracted from database constructors. We ignore the facial outline and use the interior 49 points, shown in the right side of Fig. 6.

We adopt recognition accuracy and F1-score as performance metrics. We use five-fold subject-independent cross-validation on the SPOS database and ten-fold subject-independent cross-validation on the DISFA+ database.

To compare the performance of our method to state-of-the-art research, we conduct expression distinction experiments with five methods. We use our proposed IT-RBM method, which is able to simultaneously capture global spatial patterns and complex temporal patterns. We compare it to the upper layer of the IT-RBM, which is a multi-value RBM modelling high-order spatial patterns only. The third method is HMM, a popular dynamic model capturing local temporal patterns. The fourth and fifth methods are LSTM and GRU. The first three methods are generative models, while the last two are discriminative models. The displacement of feature points are used as features for the above five methods.

For experiments with LSTM and GRU, we adopt Principal Component Analysis (PCA) to further reduce feature dimension of the landmark displacement of consecutive frames. After that, we obtain time series data with the length of T as the input of LSTM and GRU. Due to the small data size, both

TABLE 3
Results of Posed and Spontaneous Distinction Experiments

Database	DISFA+					SPOS				
Method	HMM	LSTM	GRU	RBM*	IT-RBM	HMM	LSTM	GRU	RBM*	IT-RBM
Accuracy	0.8046	0.9357	0.9442	0.9211	**0.9490**	0.7222	0.5641	0.5769	0.7735	**0.8291**
F1-score	0.8768	0.8900	0.9400	0.9095	**0.9404**	0.7619	0.6800	0.7200	0.7427	**0.8051**

**RBM is the proposed multi-value RBM.*

LSTM and GRU only have one layer of hidden units. The hidden states of the LSTM and GRU are then feed into a fully-connected network to classify expressions. For cross-validation on the DSIFA+ and SPOS databases, one fold from the training set is used as validation set for parameter selection. These are as the same as the experimental conditions used for the proposed IT-RBM. A grid search strategy is used for hyper parameters selection. Specifically, for feature dimension reduction by PCA, we get a certain number of principal components by setting different cumulative variance contribution rates ranging in $\{0.8, 0.85, 0.9, 0.95,\ 0.99, 0.999, 1\}$, for the length of input time series data, $T \in \{5, 10, 20, 30, 40, 50\}$; for the dimensions of hidden states of the LSTM (GRU), $n_h \in \{5, 10, 15, 20, 25, 30, 35, 40\}$; for the value of learning rate, $\epsilon \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$; for the size of a mini-batch, $b_n \in \{20, 40, 60, 80, 100\}$.

Since embedded spatial and temporal patterns are affected by many factors including gender and type of expression, we add two additional models using the proposed IT-RBM. One is referred to as the PS_gender model, in which we train four models: one from male posed samples, one from male spontaneous samples, one from female posed samples, and one from female spontaneous samples. The other is called the PS_exp model, in which we train a posed model and a spontaneous model for every expression type. We also examine the IT-RBM model trained on the posed and spontaneous samples, denoted as the PS model.

4.1.2 Experimental Results and Analyses

Table 3 shows the results of our experiments. From Table 3, we observe the following:

First, the proposed IT-RBM achieves superior accuracies and F1-scores than multi-value RBM; the proposed IT-RBM takes spatial patterns as well as temporal patterns into account, while the multi-value RBM only models spatial patterns. The better performance of IT-RBM demonstrates the importance of temporal patterns when distinguishing between posed and spontaneous expressions.

Second, the proposed IT-RBM achieves higher accuracies and F1-scores than HMM on both databases. As Table 3 illustrates, the accuracies of the HMM method are lower than the accuracy of IT-RBM by 0.1444 on the DISFA+ database and by 0.1069 on the SPOS database. F1-scores of the HMM method are lower than IT-RBM by 0.0636 and 0.0432, respectively. HMM is a popular dynamic model, but it can only handle three temporal relationships - precedes, follows, or equals. It is also limited to capturing local stationary dynamics because of assumptions made by the first order Markov property and stationary transition. The suggested model uses interval algebra to depict 13 complex time-point relationships, and has the ability to model global rather than only local temporal relations. This results in improved distinction between posed and spontaneous expressions.

Third, IT-RBM is superior to LSTM and GRU. Specifically, compared to LSTM, IT-RBM increases the distinction accuracies by 0.0133 and 0.2650 and F1-scores by 0.0504 and 0.1251 on the DISFA+ and SPOS databases, respectively. Compared to GRU, IT-RBM increases the accuracies by 0.0048 and 0.2522, and increases F1-scores by 0.0004 and 0.0851 on the DISFA+ and SPOS databases respectively. Although LSTM and GRU are state-of-the-art discriminative dynamic models, they are still time-slice models. Therefore, they cannot successfully represent the global and complex temporal relations between primitive facial events inherent in facial expressions, as IT-RBM does. Furthermore, recurrent neural networks achieve better performance with larger amounts of data than the used databases possess.

Fig. 7 is a graphic depiction of primitive event pairs and their corresponding temporal relations from the DISFA+ database. Fig. 7a displays the 40 selected pairs of events. Points around the eyebrow, eyelet, and lips have the most links, as these areas are crucial for expressions. Just as Ekman et al.'s research [1], [4] showed, the most telling muscles when distinguishing between expressions are orbicularis oculi and the zygomatic major. Our findings are consistent with that observation.

Figs. 7b-1 and 7b-2 illustrate temporal relations between points 20 and 29 for posed and spontaneous expressions, respectively. Figs. 7c-1 and 7c-2 are the histogram, displaying the frequencies of the 13 relations between feature point 20 and 29 in the two expressions. Figs. 7c-1 and 7c-2 show that for posed expressions, the relations of 4 and 12 occur with more frequency than the relations of 3 and 11. The inverse is true for spontaneous expressions. For relation 3 and 11, $ts_{20} - ts_{29} > 0$, which means that event v_{29} starts before event v_{20}, while for relation 4 and 12, $ts_{20} - ts_{29} < 0$, meaning that event v_{20} starts before v_{29} as shown in Table 1. This indicates that for a posed expression v_{20} starts before v_{29} in most cases, while in genuine expression v_{29} is likely to start before v_{20}. Since points 20 and 29 are representative of the right eye and the left eye respectively, we can conclude that a posed expression is more likely to begin on the right side of the face, while a genuine expression commences on the left side. This corroborates the findings of Ross and Pulusu [5].

Table 4 shows the experimental results of the PS model, the PS_gender model, and the PS_exp model. We make the following observations. First, the PS_gender model performs better than the PS model on both of the databases, with higher accuracies and F1-scores. Specifically, compared to the PS model, the PS_gender model increases recognition accuracies by 0.0134 and 0.0085 and F1-scores by 0.0152 and 0.0028 on the DISFA+ and SPOS databases respectively. This

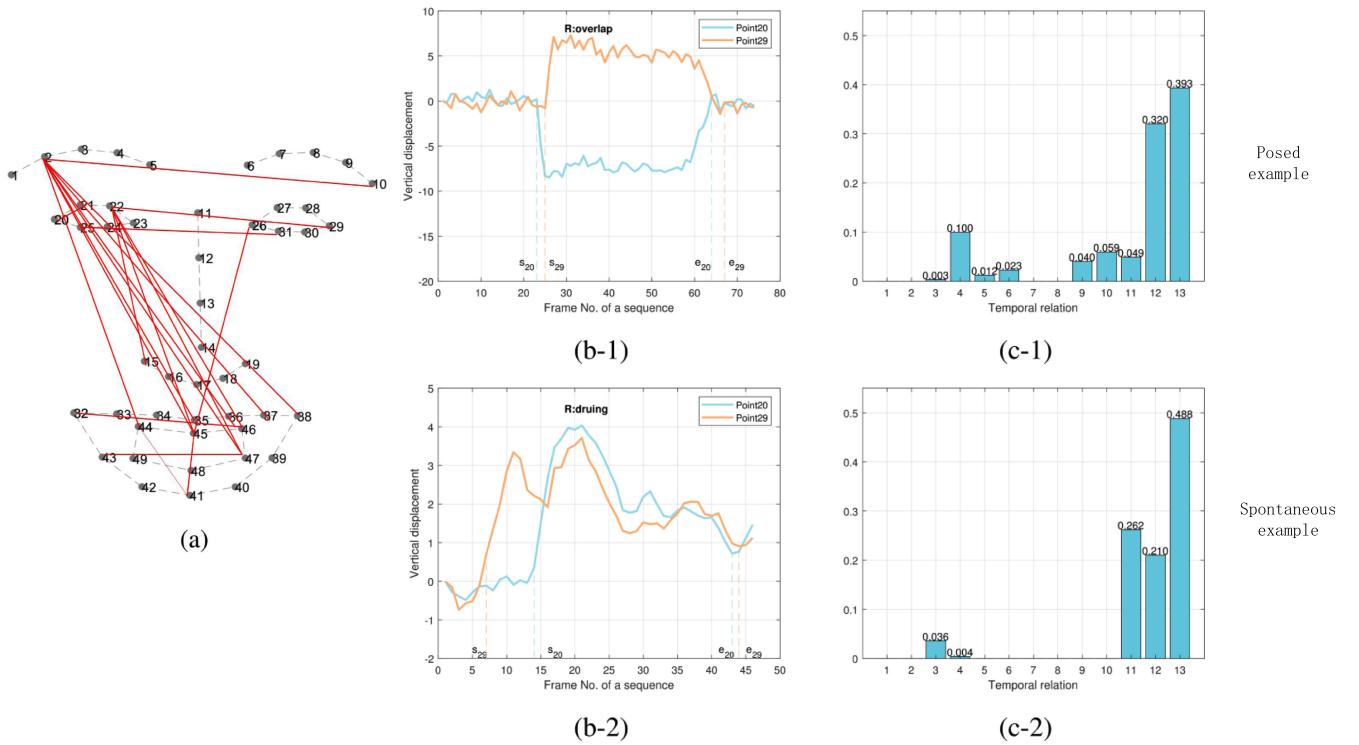


Fig. 7. (a) Graphical depiction of temporal relations selected in DISFA+. (b) Examples of relation between point 20 and 29. (c) Frequencies of thirteen relations between point 20 and point 29 with respect to posed and genuine expressions. x-axis represents the index of relationships.

indicates that gender information available only during training is useful to capture innate spatial and temporal patterns in posed and spontaneous expressions for different genders, and thus improves the distinction task.

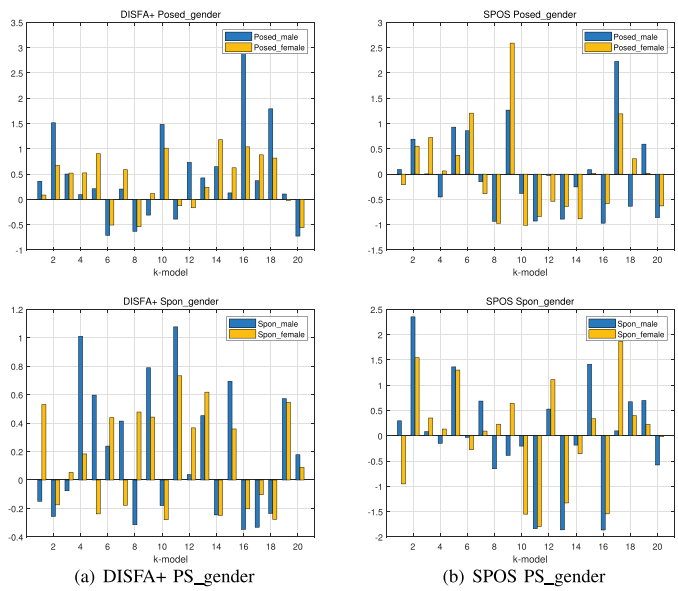
Fig. 8 graphically depicts the average weights of different movements to further analyze these spatial and temporal patterns. The x-axis represents the 20 movement patterns on the DISFA+ and SPOS databases. The y-axis represents the average value of weight w_{ik}^{j} for the Kth movement pattern. The brown bars represent the weights of PS_male model, and the blue bars are the weights of PS_female model. From Fig. 8, we can find that for some movement patterns, the weights of male and female models are either both positive or both negative, but for other movement patterns, the weight signs of male and female models are opposing. This confirms that females and males may display different spatial and temporal patterns. Therefore, the gender information available during training is beneficial for capturing more specific and precise spatial temporal patterns embedded in posed and spontaneous expressions, and results in better distinction between posed and spontaneous expressions.

Table 4 shows that the PS_exp model also performs better than the PS model, achieving superior accuracies and F1-scores in most cases. Specifically, compared to the PS model, the accuracy of the PS_exp model is 0.0025 and the F1-score is 0.0016 higher on the DISFA+ database. On the SPOS database, the accuracy improves by 0.0042 and the F1 score is 0.0042 higher. This demonstrates that the expression information available only during training is helpful for capturing inherent spatial and temporal patterns, and thus improves the performance of posed and spontaneous expression distinction.

To analyze the effect of expression type when modeling spatial and temporal patterns, we graphically depict the

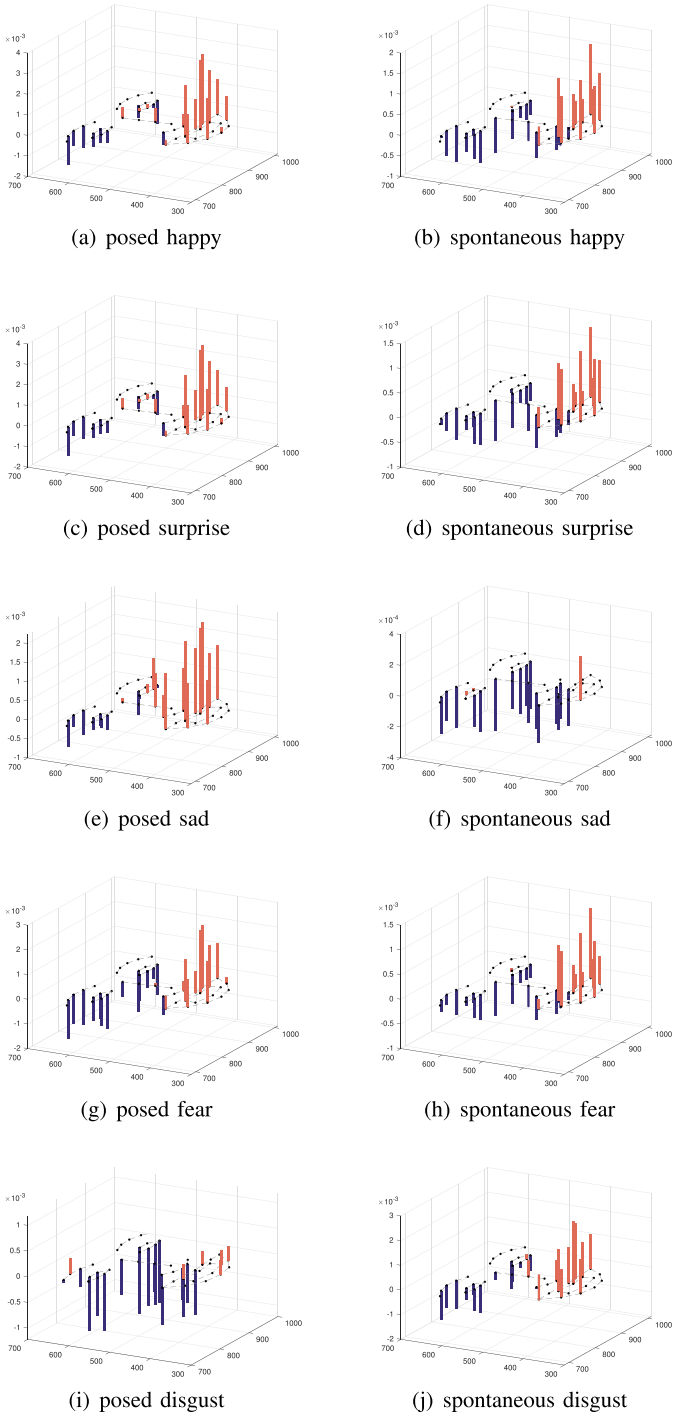
TABLE 4
Posed and Spontaneous Distinction

Database	DISFA+			SPOS		
Model	PS	PS_gender	PS_exp	PS	PS_gender	PS_exp
Accuracy	0.9490	0.9624	0.9515	0.8291	0.8376	0.8333
F1-score	0.9404	0.9556	0.9420	0.8051	0.8079	0.8093



(a) DISFA+ PS_gender (b) SPOS PS_gender

Fig. 8. The mean weight of PS_gender models at different move models on all facial points and hidden nodes. x-axis represents K move models, y-axis represents the mean value of W_{ik}^{j} at every $k = K$.



(a) posed happy (b) spontaneous happy

(c) posed surprise (d) spontaneous surprise

(e) posed sad (f) spontaneous sad

(g) posed fear (h) spontaneous fear

(i) posed disgust (j) spontaneous disgust

Fig. 9. On the DISFA+ database, the mean weight of PS_exp models at every selected facial points on a certain movement state with all hidden nodes from trained IT-RBMs. z-axis represents the mean value of W_{ij}^{k} when $k = K$ at every facial points.

average weight of the hidden nodes at every selected facial feature point for a certain movement pattern. Fig. 9 shows an example of this on the DISFA+ database. Comparing the bar graphs in the left column to those in the right, we find that there are significant differences between the weights' distribution of posed and spontaneous expressions. This indicates that the spatial patterns of posed and spontaneous expressions are absolutely different. In addition, it's clear that the W values differ significantly based on expression type; sadness is a good example. Most of the weights of the spontaneous model are negative, while most of the weights of the posed model are positive, indicating that fewer facial events are observed in spontaneous expression. Namba's [6] research shows similar results, noting that morphological properties are not observed in spontaneous facial expressions. One possible reason is that the video clip used in this study is too short for the viewer to elicit visible expressions of sadness. This explanation is supported by Eckman et al. who posit that the nature of sadness necessitates a longer-term or more personal experience [35]. For the spontaneous disgust expression, the weights in lips are positive, while not all the weights in lips are positive for posed disgust expression. Namba's [6] research showed that AU10 and AU12 were more frequently present in spontaneous disgust. Our finding is consistent with Namba's work [6], corroborating that there are expression-dependent and posed- or spontaneous-dependent differences in AUs. Adding expression information can more precisely depict the detailed patterns inherent in posed and spontaneous expressions.

TABLE 5
Comparison with Related Work of Posed and Spontaneous Expressions Distinction on SPOS

Method	accuracy
Cohn et al. [7]	0.7250
Dibeklioglu et al. [9]	0.7875
Dibeklioglu et al. [36]	0.7500
Wu et al. [37]	0.7950
Wu et al. [38]	0.8125
Wang et al. [13]	0.7479
Wang et al. [14]	0.7607
Quan et al. [16]	0.7607
IT-RBM	**0.8291**

4.1.3 Comparison with Related Work

For the task of distinguishing between posed versus spontaneous expressions, we compare our method to both model-based and feature-driven methods. Most recent model-based methods on expression distinction conducted experiments on the NVIE and the SPOS databases. As the NVIE database only provides onset and apex frames for posed expressions, it is not a viable option for the proposed IT-RBM. Instead, we used the SPOS database to compare the performance of the suggested method to current methods, as shown in Table 5. The DISFA+ database is relatively new, opening to researchers in June of 2016. Until now, no experiments on posed versus spontaneous expression distinction have been performed on this database. Therefore, we are unable to compare our work to others on this database.

From Table 5, we find that state-of-the-art model-based methods do not perform as well as the suggested IT-RBM. Compared to Wang et al. [14]'s work and Quan et al. [16]' s work, which model spatial patterns only, the IT-RBM is able to more fully represent posed and spontaneous expressions by jointly modeling the spatial and temporal patterns. This results in superior performance. The proposed method also outperforms Wu et al. [38]'s work, which is the highest-performing feature-driven method. Wu et al. [38] proposed a region-specific texture descriptor that represented local pattern changes in different areas of the face. The temporal phase of each facial region was divided by calculating the

TABLE 6
Data Distribution of CK+ and MMI

Expression	CK+	MMI
Anger(An)	45	33
Contempt(Co)	18	
Disgust(Di)	59	32
Fear(Fe)	25	28
Happy(Ha)	69	42
Sadness(Sa)	28	32
Surprise	83	41

intensity of the corresponding facial region. Then, they used a mid-level fusion strategy of SVM to combine the two feature types. By defining discriminative features, their method models the innate spatial and temporal patterns to a certain extent. However, they do not take full advantage of embedded spatial and temporal patterns as the IT-RBM does via our method's parameters and structure. Hence, the proposed method achieves superior performance.

4.2 Experiments and Analyses of Expression Recognition

4.2.1 Experimental Conditions

Expression recognition experiments are conducted on the extended Cohn-Kanade (CK+) database [39], [40] and the MMI database [41]. The CK+ database is composed of 327 posed expression samples collected from 118 subjects. It includes seven expression categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. The image sequences in this database begin at the onset frame and end with the apex frame. Therefore, only three temporal relations, i.e., before, at the same time, and after, exist in the image sequences. The MMI database is updated continuously. During April of 2017, there were 236 sequences labeled with expressions; 208 of those sequences showed the front of the face. We used these 208 image sequences from 31 subjects. There are six expression categories: anger, disgust, fear, happiness, sadness, and surprise. Table 6 shows the data distribution of the two databases. SDM is used to extract the 49 facial feature points shown on the left side of Fig. 6 [34].

Recognition accuracy is used as a performance metric. We adopt five-fold subject-independent cross-validation on the CK+ database and ten-fold subject-independent cross-validation on the MMI database.

As with posed and spontaneous expression distinction experiments, we conduct expression recognition experiments using five methods: IT-RBM, multi-value RBM, HMM, LSTM and GRU. For the experiments using HMM, the experimental results listed in [26] are directly used here. For the experiments using LSTM and GRU, the similar network structure and hyper-parameter selection strategy as those of posed and spontaneous expression distinction experiments are used.

TABLE 7
Results of Expression Categories Recognition Experiments

CK+

		An	Co	Di	Fe	Ha	Sa	Su	Acc
RBM*	An	84.44	4.44	8.89	0.00	0.00	2.22	0.00	0.8104
	Co	11.11	83.33	0.00	5.56	0.00	0.00	0.00	
	Di	13.56	0.00	76.27	1.69	0.00	5.08	3.39	
	Fe	0.00	4.00	4.00	68.00	16.00	8.00	0.00	
	Ha	0.00	1.45	0.00	1.45	97.1	0.00	0.00	
	Sa	25.00	0.00	17.86	0.00	0.00	57.14	0.00	
	Su	2.41	1.20	9.64	1.20	2.41	2.41	80.72	
		An	Co	Di	Fe	Ha	Sa	Su	
IT-RBM	An	91.11	8.89	0.00	0.00	0.00	0.00	0.00	**0.8716**
	Co	5.56	94.44	0.00	0.0	0.00	0.00	0.00	
	Di	6.78	0.00	86.44	1.69	0.00	1.69	3.39	
	Fe	0.00	4.00	0.00	72.00	16.00	8.00	0.00	
	Ha	0.00	1.45	0.00	1.45	97.10	0.00	0.00	
	Sa	17.86	0.00	3.57	0.00	0.00	78.57	0.00	
	Su	2.41	1.20	8.43	1.20	1.20	2.41	83.13	
		An	Co	Di	Fe	Ha	Sa	Su	
LSTM	An	84.44	2.22	6.67	0.00	0.00	6.67	0.00	0.8532
	Co	5.56	44.44	16.67	0.00	5.56	11.11	16.67	
	Di	5.08	1.70	81.36	1.69	1.69	1.69	6.78	
	Fe	0.00	4.00	0.00	84.00	8.00	0.00	4.00	
	Ha	0.00	1.45	0.00	4.35	94.20	0.00	0.00	
	Sa	10.71	3.57	3.57	3.57	0.00	78.57	0.00	
	Su	1.20	3.61	2.41	0.00	0.00	0.00	92.78	
		An	Co	Di	Fe	Ha	Sa	Su	
GRU	An	82.22	2.22	6.67	0.00	0.00	8.89	0.00	0.8624
	Co	0.00	50.00	0.00	0.00	0.00	11.11	38.89	
	Di	5.08	5.08	84.74	0.00	0.00	0.00	5.08	
	Fe	0.00	8.00	0.00	76.00	8.00	0.00	8.00	
	Ha	0.00	0.00	0.00	2.90	97.10	0.00	0.00	
	Sa	10.71	7.14	0.00	0.00	0.00	82.14	0.00	
	Su	0.00	4.82	0.00	0.00	1.20	1.20	92.78	
HMM [26]									0.835
ITBN [26]									0.863
Elaiwat *et al.* [42]									0.9566
Sariyanidi *et al.* [43]									0.9602

MMI

		An	Di	Fe	Ha	Sa	Su	Acc
RBM*	An	84.85	12.12	0.00	0.00	3.03	0.00	0.7740
	Di	9.38	68.75	3.13	9.38	0.00	9.38	
	Fe	0.00	3.57	71.43	10.71	3.57	10.71	
	Ha	2.38	0.00	4.76	83.33	9.52	0.00	
	Sa	9.38	0.00	3.13	3.13	81.25	3.13	
	Su	9.76	4.88	7.32	2.44	2.44	73.17	
		An	Di	Fe	Ha	Sa	Su	
IT-RBM	An	90.91	9.09	0.00	0.00	0.00	0.00	**0.8221**
	Di	6.25	81.25	0.00	3.13	0.00	9.38	
	Fe	0.00	3.57	75.00	10.71	3.57	7.14	
	Ha	2.38	0.00	7.14	83.30	7.14	0.00	
	Sa	9.38	0.00	0.00	3.13	84.38	3.13	
	Su	9.76	4.88	2.44	2.44	2.44	78.05	
		An	Di	Fe	Ha	Sa	Su	
LSTM	An	54.55	21.21	3.03	3.03	12.12	6.06	0.5769
	Di	28.13	40.63	3.13	18.75	0.00	9.38	
	Fe	3.57	10.71	28.57	14.29	14.29	28.57	
	Ha	0.00	9.52	9.52	73.81	4.76	2.38	
	Sa	21.88	9.38	25.00	3.13	40.63	0.00	
	Su	4.88	4.88	12.20	4.88	9.76	63.41	
		An	Di	Fe	Ha	Sa	Su	
GRU	An	54.55	21.21	3.03	3.03	12.12	6.06	0.5962
	Di	28.13	40.63	3.13	18.75	0.00	9.38	
	Fe	3.57	10.71	28.57	14.29	14.29	28.57	
	Ha	0.00	9.52	9.52	73.81	4.76	2.38	
	Sa	21.88	9.38	25.00	3.13	40.63	0.00	
	Su	4.88	4.88	12.20	4.88	9.76	63.41	
HMM [26]								0.515
ITBN [26]								0.597
Elaiwat *et al.* [42]								0.8163
Sariyanidi *et al.* [43]								0.7512

4.2.2 Experimental Results and Analyses

Results of our experiments on expression recognition are found in Table 7. From Table 7, we observe as follows:

First, compared to HMM [26], the accuracy of IT-RBM is higher by 0.0366 on the CK+ database and 0.3071 on the MMI database. As described in Section 2.2, HMM is a time-slice graphical model and can only capture three time-point relations. IT-RBM can not only capture 13 complex temporal relations defined by Allen's interval algebra, but also capture complex spatial patterns in facial behavior. Thus IT-RBM achieves better performance than HMM.

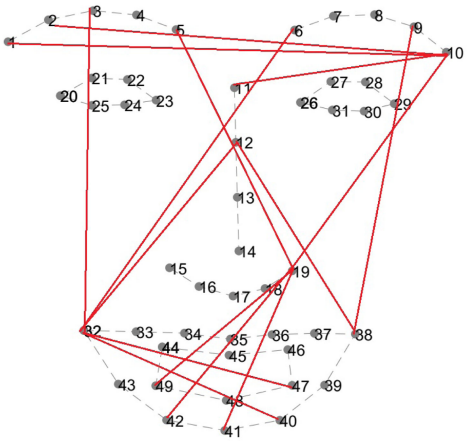


Fig. 10. Graphical depiction of the selected event pairs in the MMI database.

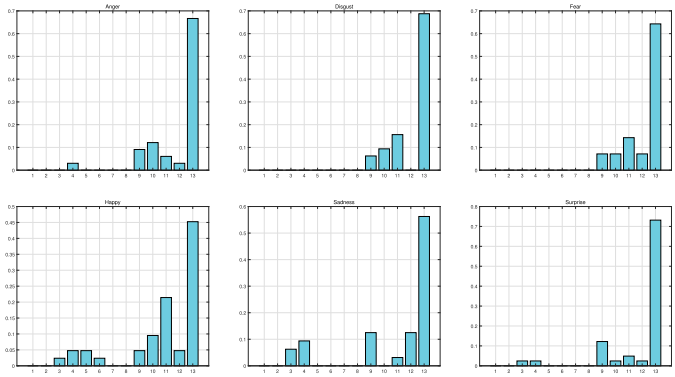


Fig. 11. Frequencies of thirteen relations among a pair of events with respect to different expressions in MMI. x-axis represents the index of relationships.

Second, the proposed IT-RBM outperforms multi-value RBM, with higher accuracies of 0.0612 on the CK+ database and 0.0481 on the MMI database. The IT-RBM not only captures global spatial relations but also temporal patterns embedded in different expressions, while the multi-value RBM is only capable of modelling inherent spatial patterns. The IT-RBM leverages the additional temporal patterns for improved expression recognition.

Lastly, the suggested method outperforms both LSTM and GRU. Compared to LSTM, IT-RBM increases the recognition accuracy by 0.0184 on the CK+ database and 0.2452 on the MMI database. Compared to GRU, IT-RBM increases the accuracy by 0.0092 and 0.2259 on the CK+ and the MMI databases, respectively. This further proves the superiority of our proposed IT-RBM in capturing and leveraging complex spatial and temporal patterns inherent in expressions for expression recognition.

In order to prove the validity of the IT-RBM for expression category recognition, Fig. 10 graphically depicts all 30 selected event pairs in the MMI database. Fig. 10 shows that the selected facial points involve all components of the face. This is reasonable, since there are 5 expressions in the database and different expressions are related to different facial muscles. Unlike Fig. 7a, in which most links appear on the left side of the face, the distribution of links in Fig. 10 is more homogeneous. This may further indicate that spatial-temporal patterns existing in posed and spontaneous expressions are not symmetrical between the left and the right sides of the face, while the spatial-temporal patterns existing in different emotion expressions are symmetrical.

Fig. 11 shows the frequencies of 13 relations between feature point 6 and point 32. From Fig. 11, we find that the frequencies of 13 interval relations among 5 expressions vary greatly. This indicates that the selected temporal relations provide discriminative information for expression recognition.

4.2.3 Comparison with Related Work

To illustrate the superiority of the proposed method IT-RBM, we compare it with the most related work (i.e., ITBN [26]) and state of the art feature-based methods [42], [43]. From Table 7, we have the following findings:

First, compared with ITBN, IT-RBM achieves the better performance on the CK+ and MMI databases. As described in Section 2.2, Although both ITBN and IT-RBM capture complex relations defined by Allen's interval algebra, ITBN uses a Bayesian network to model the local spatial patterns and IT-RBM uses a RBM to capture global spatial patterns inherent in facial behavior. Therefore, IT-RBM is more successful at capturing spatial patterns than ITBN and achieves better performance. IT-RBM can capture complex spatio-temporal patterns inherent in facial behavior, which contributes its superior performance.

Second, compared with state of the art feature-based methods, the proposed method achieves the best performance on the MMI database but the worst performance on the CK+ database. On the CK+ database, sequences begin at neutral and conclude at the peak frame. The image sequences encompass the beginning half of the expression changes only, which enforces a limit on the temporal patterns to just three relationships: A precedes B, B precedes A, and A and B commence simultaneously. Therefore, IT-RBM's ability of capturing complex temporal patterns cannot be fully displayed and IT-RBM gets the poor performance on the CK+ database. The MMI database provides the whole process of the expression change. Therefore, IT-RBM can capture whole temporal patterns and spatial patterns in the facial behavior and achieves the best performance on the MMI database. Compared with HMM, the IT-RBM get moderate improvement on the CK+ database but significant improvement on the MMI database. This is because both IT-RBM and HMM can only capture three temporal relations on the CK+ database but IT-RBM can capture more complex temporal patterns than HMM on the MMI database, which further demonstrates the importance of capturing complex temporal relations defined by Allen's interval algebra and the superior of the proposed method.

5 Conclusion

In this paper, a novel dynamic model called IT-RBM is proposed to jointly capture and leverage embedded global spatial patterns and complex temporal patterns for improved expression analysis. A facial expression is defined as a complex activity made up of sequential or temporally overlapping primitive facial events, which can further be delineated as the motion of feature points. Allen's interval algebra is

used to represent these complex temporal patterns via a two-layer Bayesian network in which the upper layer nodes represent primitive facial events, the bottom layer nodes are temporal relations between facial events, and the links between the two layers capture temporal dependencies among primitive facial events. We also suggest the use of a multi-value RBM to obtain and utilize intrinsic global spatial patterns among facial events. The visible nodes of the restricted Boltzmann machine are facial events, and the connections between hidden nodes and visible nodes model the spatial patterns inherent in expressions. In the training phase, an efficient learning algorithm is proposed to simultaneously learn spatial and temporal patterns through maximum log likelihood in the training. Samples are classified in the testing phase according to the IT-RBM with the largest likelihood. We propose an efficient inference algorithm that extends annealing importance sampling to the IT-RBM to calculate the partition function of the multi-value RBM. The results of our experiments on both expression recognition and posed and spontaneous expression distinction demonstrate that the proposed method is able to capture intrinsic facial spatial-temporal patterns, leading to superior performance compared to state-of-the-art works.

Acknowledgments

This work has been supported by the National Key R&D Program of China (2018YFB1307102) and the National Science Foundation of China (917418129).

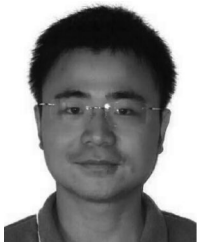
References

[1] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *J. Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.
[2] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, 2003.
[3] K. L. Schmidt, S. Bhattacharya, and R. Denlinger, "Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises," *J. Nonverbal Behavior*, vol. 33, no. 1, pp. 35–45, 2009.
[4] P. Ekman, J. C. Hager, and W. V. Friesen, "The symmetry of emotional and deliberate facial actions," *Psychophysiology*, vol. 18, no. 2, pp. 101–106, 1981.
[5] E. D. Ross and V. K. Pulusu, "Posed versus spontaneous facial expressions are modulated by opposite cerebral hemispheres," *Cortex*, vol. 49, no. 5, pp. 1280–1291, 2013.
[6] S. Namba, S. Makihara, R. S. Kabir, M. Miyatani, and T. Nakao, "Spontaneous facial expressions are different from posed facial expressions: Morphological properties and dynamic sequences," *Current Psychology*, vol. 36, pp. 593–605, 2017.
[7] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *Int. J. Wavelets Multiresolution Inf. Process.*, vol. 2, no. 02, pp. 121–132, 2004.
[8] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous versus posed facial behavior: Automatic analysis of brow actions," in *Proc. 8th Int. Conf. Multimodal Interfaces*, 2006, pp. 162–170.
[9] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 279–294, Mar. 2015.
[10] M. Seckington, "Using dynamic Bayesian networks for posed versus spontaneous facial expression recognition," Master Thesis, Dept. Comput. Sci., Delft Univ. Technol., Delft, Netherlands, 2011.
[11] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1797–1803, 2009.
[12] H. Dibeklioglu, R. Valenti, A. A. Salah, and T. Gevers, "Eyes do not lie: Spontaneous versus posed smiles," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 703–706.
[13] S. Wang, C. Wu, M. He, J. Wang, and Q. Ji, "Posed and spontaneous expression recognition through modeling their spatial patterns," *Mach. Vis. Appl.*, vol. 26, no. 2/3, pp. 219–231, 2015.
[14] S. Wang, C. Wu, and Q. Ji, "Capturing global spatial patterns for distinguishing posed and spontaneous expressions," *Comput. Vis. Image Understanding*, vol. 147, pp. 69–76, 2016.
[15] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, 2010, Art. no. 926.
[16] Q. Gan, S. Nie, S. Wang, and Q. Ji, "Differentiating between posed and spontaneous expressions with latent regression Bayesian network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4039–4045.
[17] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.
[18] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, 2017. doi: 10.1109/TAFFC.2017.2731763
[19] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
[20] T. Wu, M. S. Bartlett, and J. R. Movellan, "Facial expression recognition using gabor motion energy filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 42–47.
[21] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understanding*, vol. 91, no. 1, pp. 160–187, 2003.
[22] L. Shang and K.-P. Chan, "Nonparametric discriminant HMM and application to facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2090–2096.
[23] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst. Man Cybern., Part B (Cybern.)*, vol. 42, no. 1, pp. 28–43, Feb. 2012.
[24] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-Time Vision for Human-Computer Interaction*. Berlin, Germany: Springer, 2005, pp. 181–200.
[25] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, 2017. doi: 10.1109/TCYB.2017.2662199
[26] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3422–3429.
[27] J. Yang and S. Wang, "Capturing spatial and temporal patterns for distinguishing between posed and spontaneous expressions," in *Proc. ACM Multimedia Conf.*, 2017, pp. 469–477.
[28] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, no. 11, pp. 832–843, 1983.
[29] J. M. Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer, 2011, pp. 720–722.
[30] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[31] R. Salakhutdinov and I. Murray, "On the quantitative analysis of deep belief networks," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 872–879.
[32] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended DISFA dataset: Investigating posed and spontaneous facial expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1–8.
[33] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 868–875.
[34] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.
[35] P. Eckman, *Emotions Revealed*. Griffin, New York, USA: St. Martin's, 2003.
[36] H. Dibeklioğlu, A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 525–538.
[37] P. Wu, H. Liu, and X. Zhang, "Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 1240–1244.

[38] P. Wu, H. Liu, X. Zhang, and Y. Gao, "Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, pp. 955–967, 2017.
[39] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.
[40] 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*, 2010, pp. 94–101.
[41] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," in *Proc. 3rd Int. Workshop EMOTION (satellite of LREC): Corpora Res. Emotion Affect*, 2010, Art. no. 65.
[42] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal RBM-based model for facial expression recognition," *Pattern Recognit.*, vol. 49, pp. 152–161, 2016.
[43] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017.

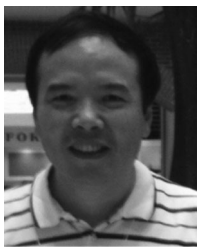
Shangfei Wang received the BS degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996, the MS degree in circuits and systems, and the PhD degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 1999 and 2002. From 2004 to 2005, she was a postdoctoral research fellow in Kyushu University, Japan. Between 2011 and 2012, she was a visiting scholar at Rensselaer Polytechnic Institute in Troy, NY. She is currently an associate professor of School of Computer Science and Technology, USTC. Her research interests cover affective computing and probabilistic graphical models. She has authored or co-authored more than 90 publications. She is a senior member of the IEEE and a member of the ACM.

Zhuangqiang Zheng received the BS degree in mathematics from Liaoning Technical University, in 2017, and he is currently working toward the MS degree in computer science at the University of Science and Technology of China, Hefei, China. His research interest is affective computing.



Shi Yin received the BS degree in automation from Central South University, in 2016, and he is now working toward the PhD degree majoring in computer science and technology at the University of Science and Technology of China, Hefei, China. His research interesting is affective computing.

Jiajia Yang received the BS degree in software engineering from Dalian Maritime University, in 2015, and she is currently working toward the MS degree in computer science at the University of Science and Technology of China, Hefei, China. Her research interesting is affective computing.



Qiang Ji received the PhD degree in electrical engineering from the University of Washington. He is currently a professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI). He recently served as a program director at the National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, the Dept. of Computer Science, University of Nevada at Reno, and the US Air Force Research Laboratory. He currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI. His research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published more than 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including NSF, NIH, DARPA, ONR, ARO, and AFOSR as well as by major companies including Honda and Boeing. He is an editor on several related the IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. He is a fellow of IAPR and the IEEE.