Multiple Emotion Tagging for Multimedia Data by Exploiting High-Order Dependencies Among Emotions
Shangfei Wang, Senior Member, IEEE, Jun Wang, Ziheng Wang, and Qiang Ji, Fellow, IEEE

Abstract—In this paper, a novel approach of multiple emotional multimedia tagging is proposed, which explicitly models the higher-order relations among emotions. First, multimedia features are extracted from the multimedia data. Second, a traditional multi-label classifier is used to obtain the measurements of the multi-emotion labels. Then, we propose a three-layer restricted Boltzmann machine (TRBM) model to capture the higher-order relations among emotion labels, as well as the relations between labels and measurements. Finally, the TRBM model is used to infer the samples' multi-emotion labels by combining the emotion measurements with the dependencies among multi-emotions. Experimental results on four databases demonstrate that our method is more effective than both feature-driven methods and current model-based methods, which capture the pairwise relations among labels by the Bayesian network (BN). Furthermore, the comparison of BN models and the proposed TRBM model verifies that the patterns captured by the latent units of TRBM contain not only all the dependencies captured by the BN but also many other dependencies that the BN cannot capture.

Index Terms—Multi-emotion recognition, multi-label classification, multimedia tagging, three-layer restricted Boltzmann machine (TRBM).

I. INTRODUCTION

RECENT years have seen a rapid increase in the size of digital multimedia collections due to the development of ubiquitous and user-friendly equipments such as smart phones and tablets. Therefore, in order to assist users in quickly finding specific data, automatic multimedia content analysis and annotation are needed to organize these collections effectively. Currently, multimedia collections, such as music, videos, and images, are not only a way to transmit knowledge and information, but also a platform to communicate within a community. Emotional tagging of multimedia collections has attracted more and more attention, since emotion is one of the key factors during communications [47], [22], [50], [18], [25], [37]. Introducing such a personal touch into multimedia content analysis could benefit both the users and businesses that create, distribute, and host the multimedia collections. For example, users could retrieve certain videos by inputting their emotional demands from a video service provider like Youtube. The video service provider could recommend videos for the target population based on emotion tag of videos.

Current research on multimedia emotion tagging mainly consists of two steps: feature extraction and classification [48]. First, several audio or visual features are extracted from multimedia, then classification methods [e.g., support vector machine (SVM)] or regression methods [e.g., Support Vector Regression (SVR)] are used to infer the multimedia data’s emotion tag as either a discrete emotional category, such as calmness, happiness or fear, or an emotional value in terms of continuous emotional dimensions, such as valence and arousal.

The assumption of most present research is that only one emotion tag can be assigned to a medium. Since these methods directly recognize labels from the extracted features without considering dependencies among labels, we refer to them as feature-driven methods. However, most multimedia data often induces a mixture of emotions in users [16]. For example, a beautiful scene may induce mixed emotions of relaxation, comfort, and happiness, and a piece of music may be characterized as both dreamy and cheerful.

Some emotions often appear together frequently, while others do not. For example, Fig. 1(a) is an image from the Memorbility database [55] which may induce mixed emotions of excitement, interest, and happiness, but it rarely induces fright. Fig. 1(b) shows several shots of a video clip from the FilmStim database [35], which conveys fear, anger, sadness, and disgust, but not tenderness or joy at all. Such phenomena of co-occurrence and mutually exclusive relationships among emotions should be considered in emotion tagging. One medium should be assigned several emotion tags simultaneously. Thus, emotion tagging should be formulated as a multi-label classification problem.

Presently, few researchers regard emotion tagging of multimedia as a multi-label classification problem, except for a small number of studies on emotion recognition from music...
or video data [48]. Current multi-label classification methods which address label dependencies directly either ignore the label correlations or fix the relations as a pairwise or a subset of label combinations present in the training data. They cannot effectively explore the co-occurrence or mutually exclusive relationships among emotional labels. Current multi-label classification methods, which address label dependencies indirectly through features and hypothesis, are usually limited to certain learning methods or have higher computation costs. Recently, we [48] directly exploit the dependencies among emotion tags by a Bayesian network (BN), where each node represents an emotion label and the links and conditional probabilities capture the probabilistic dependencies among multiple emotions. Our work successfully captures the co-occurrence and mutual exclusion relationships among emotional labels. Any machine learning method can be adopted to obtain label measurements. Using the learned Bayesian Network, the true labels are inferred by combining the relationships among labels with the labels’ estimates obtained from a current machine learning method. We refer to it as a pairwise model-based emotion tagging method. However, due to the first-order Markov assumption of BN, this model can only capture the pairwise dependencies of the labels instead of the higher-order relations among the multi-labels. Furthermore, finding the optimal structure of a large emotion label network for multimedia is difficult. An advanced multiple emotion recognition method from multimedia should capture not only pairwise but also higher-order dependencies among emotions.

Unlike a regular BN, the Restricted Boltzmann Machine (RBM) and its variants can model higher-order dependencies among random variables by introducing a layer of latent units. Therefore, we further proposed a Three-Layer RBM (TRBM) model to exploit the higher-order relationships among labels for action unit recognition [51]. In this paper, we propose an approach using RBM to model and enhance the complex joint distributions over structured multiple emotion labels for multiple emotion tagging of multimedia data. Each visible node represents an emotion label, and weights among the visible nodes and hidden nodes capture the higher-order dependencies among multiple emotions. Our approach consists of several steps: first, a traditional multi-label classifier is adopted to obtain the measurements of the emotion tags from the audio-visual content. Then, a TRBM is automatically constructed to model the higher-order dependencies among emotion tags. The constructed TRBM is employed to infer the true tags for a medium from their measurements. We conduct experiments on four benchmark databases: the Music Emotion database, the NVIE video database (a multiple emotion video database), the Memorability database, and the FilmStim database. Experimental results show that our method can successfully exploit the higher-order relationships among emotions, and thus improving the performance of traditional multi-label classifiers or the method in which a BN is used to capture the pairwise relations among multi-labels. By exploiting RBM’s full connections between latent and visible nodes, our model allows capturing dependencies among all emotions. The captured additional relationships beyond the pairwise relationships offer additional constraints on the emotion tags and hence can further improve the performance as shown by the experiments.

The rest of the paper is organized as follows: Section II presents an overview of the related works. The detailed introduction of our method is given in Section III. Section IV discusses the experimental results. Finally, the paper is concluded in Section V.

II. RELATED WORK

A. Multi-Label Classifications

Multi-label classification is the classification problem in which one sample can be assigned to more than one target
label simultaneously. Successfully exploring the higher-order dependency inherent in multiple labels is the key to facilitating the learning process. Current studies either directly or indirectly consider dependencies among labels. The former explore label dependencies directly from the target labels, and the latter consider label dependencies with the help of features or hypotheses.

One special case of the first group is Binary Relevance (BR), which decomposes multi-label problems into multiple independent binary classification problems by using one label for one classification, ignoring the correlations among labels. Other studies that directly consider dependencies among labels address the pairwise relations between labels (such as Calibrated Label Ranking (CLR) [14]), or the fixed label combinations present in training data (such as Label Powerset (LP) [7]), or a random subset of the combinations (such as Random k-labelsets (RAKEL) [44]). Since the relations among labels may be beyond pairwise, and cannot be expressed by a fixed subset of labels existing in training data, the first group may not capture the label relations effectively.

The second group considers label dependencies with the help of features or hypotheses. Some researchers expand current machine learning methods to handle multi-label classification by modifying the hypothesis. For example, Zhang et al. [65] proposed multi-label back-propagation (BPMLL), which defines a new error function to capture the characteristics of multi-label learning for a back-propagation algorithm, i.e., the labels belonging to an instance should be ranked higher than those not belonging to that instance. Inspired by the k-Nearest Neighbors (kNN) algorithm, Zhang et al. [66] proposed multi-label k-nearest neighbor (MLKNN), a multi-label lazy learning approach. The nearest neighbors in the training set are identified for each unseen instance. Then, based on statistical information gained from the label sets of these neighboring instances, a maximum posteriori principle is utilized to determine the label set for the unseen instance. Other researchers use labels as additional features. For instance, Godbole and Sarawagi [15] adopt a two-stage classification process by stacking the outputs of BR along with the full original feature space into a separate meta-classifier. Read et al. [34] link n classifier into a chain, creating a classifier chain model, where the feature space of each classifier in the chain is extended with the label associations of all previous classifiers. Zhang and Zhang [64] propose a Bayesian network to model the dependencies among label errors, and construct a binary classifier combining the features and the parental labels, which were regarded as additional features. Huang et al. [19] propose to model the label relations by a hypothesis reuse process. When the classifier of a certain label is learned, all trained hypotheses generated for other labels are taken into account via weighted combinations. With the help of features and hypotheses, these methods can model the flexible dependencies among labels to some extent, but their computational costs are usually much higher compared to methods that directly model the dependencies among labels.

### B. Emotion Tagging of Multimedia

Assigning emotion tags to multimedia data has been an active research area in recent decades [47], [22], [50], [18], [25], [37]. Although music, images, and videos are different modalities, the research regarding emotion tagging of these three media obeys a similar framework. First, emotional descriptors are adopted to capture the users’ subjective tagging of the emotional content of multimedia. Audio or visual features are then extracted from multimedia. After that, classification methods or regression methods are used to assign an emotion tag or a point in emotional dimension space [48]. In this section, we use video as an example to give a brief review of emotion tagging of multimedia. Similar emotional descriptors and machine learning methods are used for images or music pieces. A comprehensive overview on emotion tagging of music pieces, images and videos can be found in [47], [22], [50], [18], [25], and [37].

Two kinds of emotional descriptors have been proposed to capture affective content of videos: a categorical approach, and a dimensional approach. The most frequently used categories in the field of video emotional content analysis are Ekman’s six basic emotions [12], including happiness, sadness, anger, disgust, fear, and surprise [23], [24], [39], [28], [57], [6], [20], [46], [32], [62]. Others categories, such as amusement [2], [1], boredom [70], excitement [52], horror [59], [10], [29], [30], [70] or funny [13], have also been used to describe emotional content of videos. For the dimensional approach, terms such as valence-arousal-dominance are frequently used [36], [11], [3], [45], [40], [17]. In addition, a few work adopt other dimensions, such as natural, temporal, and energetic dimensions [4].

Furthermore, the categorical and dimensional definitions of emotion are related. For example, in Russell’s affective model, happiness always belongs to the first quadrant of the model, and anger clearly belong to the fourth quadrant. Some studies use the relation between categorical and dimensional definitions for emotional category tagging from videos. For example, Arfin and Cheung [2], [1] regress video features to the pleasure-arousal-dominance space first, and then translate the emotional dimensions into emotion categories.

Video content can be captured by various visual and audio features. Specifically, various color, motion, and lighting features are extracted from visual images, and scripts are extracted from print and other graphics channels to characterize videos’ visual content. For audio data, continuous prosodic features like speech energy, pitch, and fundamental frequency as well as spectral features like Mel-frequency Cepstral Coefficient (MFCC) are the most commonly used speech features. Dynamics, timbre, harmony, and rhythm are widely used music features. Many audio features are shared among different audio types (i.e. speech, music, and background sound), including energy-related features, spectral centroid, spectral rolloff, and spectral flux.

To model the mapping between the video features and the discrete emotional descriptor, many machine learning methods have been investigated, such as support vector machine (SVM) [53], multi-layer feed-forward neural networks (NNs) [52], AdaBoost [23], Gaussian mixture models (GMMs) [63], k-Nearest Neighbor (kNN) [61], hidden Markov models (HMMs) [39], [58], dynamic Bayesian networks (DBNs) [3], and conditional random fields (CRFs) [60]. Wang et al. [46] adopt a specially adapted variant of SVM to classify films into anger, sadness, fear, joy, surprise, and neutral. Watanapa et al. [52] propose a two-stage sieving artificial neural network to classify movie clips into excitement, joy, and sadness. They first specializes in filtering the excitement class, and then classifies joy and sadness. Yazdani et al. [63] adopt GMM to analyze affective
content from music video clips. Yazdani et al. [61] classify music video clips into high, low, and neutral arousal, or positive, neutral, and negative valence using KNN. Kang [23] proposes two HMMs of different topologies to map low-level audio-visual features to high-level emotional states (i.e., fear, sadness, joy, and normal state). Xu et al. [56] develop a four-state HMM to classify audio emotional events such as laughing or horror sounds in comedy and horror videos. Ariffin and Cheung [2], [1] construct an n-level dynamic Bayesian network for affective classification in the pleasure-arousal-dominance space from video features, and then translate the emotional dimensions into emotion categories. Teixeira et al. [28] propose a hidden Markov model and an autoregressive hidden Markov model to estimate the values of pleasure, arousal, and dominance for each video segment, and then translate the resulting pleasure-arousal-dominance values into emotion categories.

Other than a classifier, a regressor is used to map the features to the continuous emotional dimensions. Two kind of approaches are adopted for such mapping. One is to manually define the mapping function between low-level features and dimensional emotional descriptors, and the other is to use general regression, such as polynomial regression [8], neural network [52] or support vector regression [4], [5], [9], [8], [68], [67], [69] to learn the mapping functions from data. For example, Hanjalic and Xu [17] manually define an analytic time-dependent function to map the motion intensity, cut density, and sound energy onto the arousal dimension, and the pitch-average to the pleasure dimension. Zhang et al. [68], [67] adopt support vector regression to learn the mapping from motion intensity, short switch rate, zero crossing rate, tempo, and beat strength to arousal dimension, and that from lighting, saturation, color energy, rhythm regularity, and pitch to the valence dimension. Canini et al. [4], [5] compare three regressive procedures, i.e., polynomial combination, feed-forward neural network trained by a back-propagation algorithm, and support vector regression, to predict the natural, temporal and energetic dimensions from several audio and visual features.

To the best of our knowledge, most present research of emotion tagging of images or video assumes that for each image or video, there is only one emotion tag or one point in emotional dimensional space. However, the disjointedness of the labels is not valid in emotion detection from images or videos. Gross et al. ’s study [16] indicate that the videos that induce anger may also induce some degree of disgust, sadness, fear, and surprise. However, the videos that induce anger and disgust are unlikely induce high levels of happiness. The phenomena of complex co-occurrence, and mutual exclusion for emotional categories are also revealed in [33]. There has been little research considering multiple emotion tagging of images and videos until now [48].

There are a small number of studies that consider assigning multiple emotion labels to a music piece. Li and Ogihara [27] may be the first to realize that emotion detection from music should be formulated as a multi-label classification problem. They decompose the problem into a set of binary classification problems, and adopt SVM as the classifier for each binary classification. Their experiment uses a collection of 499 sound files created from 128 music albums and labeled with ten adjective groups of Farnsworth. Wieczorkowska et al. [54] regard samples with many labels as positive examples for each class corresponding to the labels. Their testing data set contains 875 samples, annotated with 13 labels. Trohidis et al. [41] compare four multi-label classification algorithms, i.e., binary relevance (BR), label powerset (LP), random k-labelsets (RAKEL), and multi-label k-nearest neighbor (MLkNN), for emotion detection from music. They further construct the only publicly available music emotion database, including 593 songs with six emotion labels. Later, Trohidis et al. [42] extend their study by comparing seven algorithms, including BR, LP, RAKE, ranking by pairwise comparison (RPC), calibrated label ranking (CLR), multi-label back-propagation (BPMLL), and MLkNN.

All the above works demonstrate that multi-label modeling is successful at emotion detection from music. The methods used in these studies can be clarified into two groups: those which consider the label dependencies from the target labels, and those which explore label dependencies indirectly with the help of hypothesis. The former includes BR, LP, RAKE, and CLR. They either ignore the label correlations, such as BR, or fix the relations as a pairwise or subset labels combination in the training data, such as CLR, LP, RPC and RAKE. Thus, they cannot effectively exploit the co-occurrence or mutually exclusive relations among emotions. BPMLL and MLkNN explore label dependencies indirectly with the help of hypothesis. They respectively extend back-propagation and k-nearest neighbor to handle multi-label data directly. With the modified hypothesis, they can model the flexible label dependencies to some extent. However, their computation costs are usually higher than other methods, especially when the feature dimension and sample size are large. In addition, the first group transfers multi-classification tasks into several single tasks by directly exploring the dependence from target labels. Thus, any machine learning method can be used. The second group is limited to specific learning methods by modified hypothesis. Therefore, we prefer to model the higher-order dependencies from the target label directly.

Recently, we [48] proposed a framework of multi-label multimedia emotion tagging which captures the relations among multi-labels to enhance multi-label recognition. This framework is applicable to emotion tagging of music pieces as well as images and videos. A BN is automatically constructed to systematically capture the dependencies among emotion tags. The nodes of the BN represent the labels. The links and their parameters capture the probabilistic relations among labels. The emotion relationships encoded in a BN are more flexible than the fixed subset of relationships used in existing work. Nevertheless, due to the Markov assumption, BN is limited at capturing the relationships between pairs of emotions such as co-occurrence, co-absence, and mutual exclusion. In addition, it is difficult to obtain the optimal structure, especially with a large number of emotions. By introducing latent nodes, restricted Boltzmann machine can capture not only pairwise but also higher-order dependencies among labels. Thus, we further propose a Three-Layer RBM model to exploit the higher-order relationships among labels for action unit recognition [51]. In this paper, we extend our TRBM model to multi-label emotion tagging of multimedia. Our model consists of three layers. The top two layers are similar to RBM, with a visible layer and a latent layer. The visible nodes represent emotion labels, and the weights between visible nodes and hidden nodes capture the higher-order dependencies of multiple emotion labels. The
bottom layer represents the measurements and the the links between the second and the third layer captures the relations between emotion labels and their measurements.

III. MULTIPLE EMOTION TAGGING METHODS

The framework of our approach, shown in Fig. 2, consists of three steps: feature extraction, measurement acquisition, and multi-emotion relationship modeling by TRBM. The training phase of our approach involves training a traditional multi-label classification method for emotion tag measurement acquisition, and training the TRBM to capture the semantic relationships among emotion tags. For measurement acquisition, we employ audio and visual features to represent the media, and perform preliminary classification using a traditional multi-label algorithm. Given the measurements, we infer the emotion tags of media through a probabilistic inference with the TRBM model. The details are provided as follows.

A. Feature Extraction

Four benchmark databases are used in the work, including the Music Emotion database [41], the NVIE video database (a multiple emotion video database collected in [48]), the Memorability database [21] and the FilmStim database [35]. Therefore, we only focus on the music features, video features, and image features.

Due to copyright issues, the Music Emotion database [41] does not provide the original music clips, but it does provide 8 rhythmic features and 64 timbre features of each sample. The rhythmic features are derived by extracting periodic changes from a beat histogram. The timbre features consist of the first 13 MFCCs, spectral centroid, spectral rolloff and spectral flux and their means, standard deviations, mean standard deviations, and standard deviations of standard deviation over all frames. We adopt these features in the experiments for the Music Emotion database.

The constructors of the NVIE video database and the FilmStim database do not provide features. Thus, We extract visual and audio features from these two video databases. For visual features, lighting, color, and motion are powerful tools to establish the mood of a scene and affect the emotions of the viewer according to cinematography and psychology. Thus, three features, named lighting key, color energy, and visual excitement are extracted from video clips. 31 audio features widely used in the video tagging field [26] are extracted, including average energy, average energy intensity, spectrum flux, Zero Crossing Rate (ZCR), standard deviation of ZCR, 13 MFCCs, log energy of MFCCs, and the standard deviations of the above 13 MFCCs. The features are averaged over the whole clip. Therefore, a total of 34 features are acquired to represent each video signal. These visual and audio features are complementary for emotion tagging of videos. The details of these features can be found in [49].

The Memorability database [55] provides several features, including GIST feature, HoG features, dense SIFT features, sparse SIFT histograms, SSIM features, tiny images, line features, Texton histograms, color histograms, geometric probability map and geometry-specific histograms. These features are used for the experiments on the Memorability database.

B. Emotion Tag Measurement Acquisition

Any commonly used multi-label classification method can be adopted to obtain the measurements of the emotion tags. Here we select binary relevance (BR) [43], which transforms the multi-label classification task into several single-label classification tasks. Let $D = \{ (x_i, \lambda_i) \}_{i=1}^{N}$ represent the training data, where $x_i \in \mathbb{R}^d$ is the feature vector of the $i$th sample and $N$ is the number of training samples. $\lambda_i$ is the multiple target labels, which is an $n$-dimensional vector. $\lambda_{ij}$ represents the $j$th label of the $i$th label, and $n$ is the number of labels.

BR is the most widely used problem transformation method. It considers each label independently. First, it changes original data set to $n$ data sets, each data set $D_j$ for one label $\lambda_j$. Then, any traditional classification algorithm can be used to obtain the classifiers $h_j$ using $D_j$. For a new instance $X_i$, each classifier $h_j$ outputs a binary label $Z_{ij}^{h} = h_j(X_i)$. Then, the combination of the labels predicted by $n$ classifiers $(\cup_{j=1}^{n} Z_{ij})$ is adopted as the final output of sample $X_i$. Support vector machine is used

---

**Fig. 2.** Framework of our proposed multi-label recognition approach.
as the classifier. BR assumes that the labels are independent, ignoring the correlations among those labels.

The outputs of the above method are binary vectors, indicating whether a music piece, video, or image has a certain emotion tag or not. The binary vector is used as the measurement for the TRBM model in the following step.

C. Higher-Order Dependencies Modeled by Three-Layer Restricted Boltzmann Machine

The proposed three-layer RBM model is shown in Fig. 3, similar to the hybrid RBM proposed in [51] for AU recognition. The middle layer contains the binary visible units \( \{ \lambda_1, \ldots, \lambda_n \} \), representing the state of emotion tags. A layer of binary latent units \( \{ h_1, \ldots, h_m \} \) are imposed upon the labels. Each latent unit is connected to all the emotion tags, and therefore is used to model their higher-order relationships. The variables \( \{ \lambda_1, \ldots, \lambda_n \} \) in the bottom layer stand for the emotion tag measurements obtained as outlined in Section III-B. They are attached to each multi-label variable, providing preliminary recognition results using a traditional multi-label classification method.

Equation (1) defines the total energy of the TRBM model

\[
E(\lambda, \lambda; h; \theta) = - \sum_i \sum_j \lambda_i W_{i,j} h_j - \sum_j c_j h_j - \sum_i b_i \lambda_i - \sum_i \sum_j W_{i,\lambda_i} \lambda_i \lambda_i
\]

where \( \{ b_i \} \) and \( \{ c_j \} \) represent the biases of the emotion tag nodes and latent nodes. The probabilistic relations between each pair of emotion tag \( \lambda_i \) and latent unit \( h_j \) are captured by the first set of parameters \( W_{i,j} \), and the second set of parameters \( W_{i,\lambda_i} \) captures the probabilistic relation between each pair of emotion tag measurement \( \lambda_i \) and the emotion tag \( \lambda_i \).

The TRBM model shown in Fig. 3 can be decomposed into two parts: the upper part, consisting of the top two layers, which captures the relations among multiple emotion labels, and the lower part, consisting of the bottom two layers, which captures the relations between emotion tag measurements and true labels.

The upper part of the model can be considered a traditional RBM model, encoding the high-level semantic relationships among multiple emotion labels through the connections from each latent unit to all visible units. The parameters \( W_{i,j} \) implicitly describe the captured multi-label relationships. Considering the \( j^{th} \) latent unit \( h_j, j = 1, 2, \ldots, m \), its probabilistic dependency with each emotion tag is measured by the pairwise energy \( E(\lambda_i, h_j) = -W_{i,j} \lambda_i h_j \), \( i = 1, \ldots, n \). Obviously, a larger \( W_{i,j} \) indicates lower energy, and thus higher probability. Therefore, the larger \( W_{i,j} \) is, the more likely \( \lambda_i \) will be present. Conversely, the smaller \( W_{i,j} \) is, the more likely \( \lambda_i \) will be absent. To be specific, model parameter \( W_{i,j} \) captures a specific presence and absence pattern of the multiple emotion tags. For example, the parameter \( W_{i,j} \) for \( h_j \) in Fig. 3 graphically depicts one of the patterns. In this case, \( h_j \) captures a pattern where \( \lambda_1, \lambda_2 \) are very likely to be absent, \( \lambda_1, \lambda_2 \) are very likely to occur, and \( \lambda_3, \lambda_4 \) tend to occur yet are less likely than \( \lambda_5, \lambda_6 \).

The lower part of the TRBM model is essentially equivalent to a set of linear classifiers weighted by \( W_{i,h} \), which integrate multiple emotion labels and emotion tag measurements obtained by a traditional multi-label classification method. The energy between the measurements and the corresponding emotion label is defined as the energy \( E(\lambda_i, \lambda_i) = -\sum_i W_{i,h_i} \lambda_i \lambda_i \).

As a whole, combining the two parts, the proposed TRBM model captures the information from both the emotion tag measurements and the higher-order relationships among multiple emotions.

1) Model Learning: Given the training data \( D = \{ \lambda_i, \lambda_i \}_{i=1}^N \), \( \lambda_i \) represents the ground truth multi-labels of the \( i^{th} \) training sample and \( \hat{\lambda}_i \) represents the measurements. Parameters, including \( b, c, W_{i,\lambda} \) and \( W_{i,h} \), are learned in a discriminative manner by maximizing the log conditional likelihood, as shown in (2)

\[
\theta^* = \arg\max \log P(\lambda_i, \hat{\lambda}_i, \theta)
\]

\[
- \arg\max_\theta \log \left( \log \sum_{\lambda_i} e^{-E(\lambda, \lambda, \lambda_i, \theta)} - \log \sum_{\lambda_i} e^{-E(\lambda, \hat{\lambda}, \lambda_i, \theta)} \right)
\]

where \( \theta \) consists of all the parameters of the model, and is maximized with the gradient descent method in which the gradient can be calculated with (3)

\[
\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \log \sum_h e^{-E(\lambda, \lambda, h, \theta)} - \frac{\partial}{\partial \theta} \log \sum_{\lambda_i} e^{-E(\lambda, \lambda, \lambda_i, \theta)}
\]

\[
= - \sum_h \frac{1}{e^{-E(\lambda, \lambda, h, \theta)}} \sum_{\lambda_i} \left( e^{-E(\lambda, \lambda, h, \theta)} \frac{\partial E(\lambda, \lambda, h, \theta)}{\partial \theta} \right)
\]

\[
+ \sum_{\lambda_i} \frac{1}{e^{-E(\lambda, \lambda, h, \theta)}} \sum_{\lambda_i} \left( e^{-E(\lambda, \lambda, h, \theta)} \frac{\partial E(\lambda, \lambda, h, \theta)}{\partial \theta} \right)
\]

\[
= \sum_h \left( e^{-E(\lambda, \lambda, h, \theta)} \frac{\partial E(\lambda, \lambda, h, \theta)}{\partial \theta} \right)
\]

\[
+ \sum_{\lambda_i} \left( e^{-E(\lambda, \lambda, \lambda_i, \theta)} \frac{\partial E(\lambda, \lambda, \lambda_i, \theta)}{\partial \theta} \right)
\]

\[
= \left( \frac{\partial E(\lambda, \lambda, h, \theta)}{\partial \theta} \right)_{p(h|\lambda, \lambda, \theta)} + \left( \frac{\partial E(\lambda, \lambda, \lambda_i, \theta)}{\partial \theta} \right)_{p(\lambda, \lambda_i, \theta)}
\]
Calculating the gradient involves inferring \( P(h|\lambda, \hat{\lambda}, \theta) \) and \( P(h|\lambda, \hat{\lambda}, \theta) \). Calculating \( P(h|\lambda, \hat{\lambda}, \theta) \) can be calculated with (4), where \( \sigma(x) = 1 / (1 + e^{-x}) \) is the sigmoid function. Unfortunately, \( p(h, \lambda|\lambda, \theta) \), similar to \( p(h, \lambda|\theta) \) in RBM, is intractable to compute, and contrastive divergence (CD) algorithm [31] is proposed to solve this problem, where a one-step Gibbs sampling from the data is performed to approximate the parameter.

Therefore, we extend the CD algorithm to learn the TRBM model in a similar way. The basic idea is to approximate \( p(h, \lambda|\lambda, \theta) \) by sampling \( h \) with (4) and then sampling \( \lambda \) with (5). The detailed algorithm for learning the parameters is shown in Algorithm III-C.1, where \( \eta \) is the learning rate defined manually, and \( D^+ \) and \( D^- \) separately represent the positive term and negative term in (3).

\[
P(h_j|\lambda, \hat{\lambda}) = \sigma \left( c_j + \sum_i W^h_{ji} \lambda_i \right) \quad (4)
\]
\[
P(\lambda_i|h, \hat{\lambda}) = \sigma \left( b_i + \sum_j W^l_{ij} h_j + \sum_t W^l_{it} \hat{\lambda}_t \right) \quad (5)
\]

Algorithm 1 Revised contrastive divergence algorithm for learning the proposed model [51]

1: Input: Training data \( \{\lambda_i \in R^{1 \times d}, \hat{\lambda}_i \in R^{1 \times n}\}^N_{i=1} \).
2: Output: Model parameters \( W^1 \in R^{n \times m} \).

repeat
3: for each training instance \( (\hat{\lambda}, \lambda) \)
4: \( D^+_2 = \lambda \hat{\lambda} \)
5: Sample \( h^+ \sim P(h|\lambda, \hat{\lambda}) \) with (4)
6: Calculate the positive gradient \( D^+_1 = \lambda h^+ \)
7: Sample \( \lambda^- \sim P(\lambda|h^+, \hat{\lambda}) \) with Eq. (5)
8: \( D^-_2 = \lambda^- \hat{\lambda} \)
9: Sample \( h^- \sim P(h|\lambda^-, \hat{\lambda}) \) with Eq. (4)
10: Calculate the negative gradient \( D^-_1 = \lambda^- h^- \)
11: Update:
\[ W^1 = W^1 + \eta(D^+_1 - D^-_1), b = b + \eta(\lambda - \lambda^-) \\
W^2 = W^2 + \eta(D^+_2 - D^-_2), c = c + \eta(h^+ - h^-) \]
12: Endfor
until Converges

2) Inference: Given a testing sample where the acquired measurement is \( \hat{\lambda} \), then each emotion tag \( \lambda_i \) can be calculated by maximizing its posterior probability given \( \hat{\lambda} \) with (6)

\[
\lambda^*_i = \arg\max_\lambda P(\lambda_i|\hat{\lambda}) \quad (6)
\]

Computing \( P(\lambda_i|\hat{\lambda}) \) requires marginalizing over all the latent variables \( \{h_j\}^m_{j=1} \) and other labels \( \{\lambda_k\}^m_{k \neq i} \), and thus could be intractable. \( P(\lambda_i|\hat{\lambda}) \) can be efficiently computed with the Gibbs sampling method by iteratively sampling \( h \) from \( P(h|\lambda, \hat{\lambda}) \) and sampling \( \lambda \) from \( P(\lambda|h, \hat{\lambda}) \). Sampled instances of each \( \lambda_i \) are used to calculate the corresponding marginal probability. Algorithm III-C.2 gives detailed information.

IV. EXPERIMENT

A. Experimental Conditions

Until now, there have only been a few multimedia databases which contain multiple emotion labels. In this work, experiments are performed on four databases, including the Music Emotion database [41], the NVIE video database [48], the Memorability database [21], and the FilmStim database [35].

The Music Emotion database contains 593 songs categorized into one or more of the following classes of emotions: amazed-surprised (amazed), happy-pleased (happy), relaxing-calm (relaxing), quiet-still (quiet), sad-lonely (sad), and angry-fearful (angry). The duration of each music clip is 30 seconds, and the frame rate of speech is 22.05 kHz. Detailed information about the database can be found in [41].

The NVIE video database [48] consists of 72 videos, and the lengths of the videos vary from thirty seconds to five minutes. The frame rates of the speech and video are 44 kHz and 30 fps. For each video shot, emotional valence and arousal ranging from -2 to 2 are labeled, as well as six basic emotional categories ranging from 0 to 4, where 0 indicates no particular feeling and 4 indicates a strong feeling. A threshold is set to transform the intensity of emotion tag to a binary tag, which represents if a certain emotion is present or not. If the intensity is larger than the threshold, the tag is set to 1; otherwise, it is 0. The threshold of emotional categories is 0.2, and the threshold for valence and arousal is 1. Detailed information about the database can be found in [48].

The Memorability database consists of 2222 natural images of everyday scenes and events selected from the SUN database [55], as well as 923 attributes for each image. There are multiple emotion labels for each image, i.e., frightening, arousing, funny, engaging, peaceful, exciting, interesting, mysterious, strange, striking, makes you happy, and makes you sad.

The FilmStim database is collected by Schaefer et al. [35], and includes a total of 64 videos ranging from one to seven minutes in length. Six emotional discreetness scores (for anger, disgust, sadness, fear, amusement, and tenderness) are computed for each video. The multiple emotion labels are all transformed to binary according to the mean score of each label. If the score is larger than or equal to the mean value, it is set to be 1; otherwise, it is set to be 0. The sample distribution of these four databases is presented in Table I.
To validate the proposed model, three kinds of experiments are conducted: emotion tagging using current feature-driven methods, which is the same as the measurement acquisition; emotion tagging using current model-based methods, which adopt BN to model pairwise label dependencies; and our proposed method. Furthermore, we compare our method with current multi-label methods, i.e., BPMLL and MLkNN.

In this work, model selection is adopted to determine the number of hidden nodes and the parameters of the Three-Layer RBM model, and 10-fold cross-validation is performed to verify the effectiveness of our method.

### B. Evaluation Metrics

The evaluation metric of multi-label classification is different from that of single label classification, since for each instance there are multiple labels which may be classified partly correct or partly incorrect. Thus, there are two kinds of commonly used metrics, example-based and label-based measures [38], evaluating the multi-label emotion tagging performance from the view of instances and labels respectively. We adopt both measures in this work. Let \( Y_i^j \) denote the \( j \)-th true label for instance \( i \), which is a binary value; and \( Z_i^j \) is the \( j \)-th predicted label for instances \( i \); \( N \) represents the number of the instances; and \( n \) is the number of labels. The example-based measures (accuracy, recall and F1-measure) are separately defined in (7)–(9), and the label-based measures (micro averaging recall, micro averaging F1-measure, macro averaging recall, and macro averaging F1-measure) are separately defined in (10)–(13).

\[
\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i \cap Z_i}{Y_i \cup Z_i} \quad (7)
\]

\[
\text{Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap Z_i|}{|Y_i|} \quad (8)
\]

\[
F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{2 Y_i \cap Z_i}{|Y_i| + |Z_i|} \quad (9)
\]

\[
\text{Recall, } R_{\text{micro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{n} I_{Y_i^j} I_{Z_i^j}}{\sum_{j=1}^{n} I_{Y_i^j}} \quad (10)
\]

\[
F_{1-\text{micro}} = \frac{2 \sum_{j=1}^{n} \sum_{i=1}^{N} Y_i^j Z_i^j}{\sum_{j=1}^{n} \sum_{i=1}^{N} Y_i^j + \sum_{j=1}^{n} \sum_{i=1}^{N} Z_i^j} \quad (11)
\]

\[
\text{Recall, } R_{\text{macro}} = \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{i=1}^{N} Y_i^j Z_i^j}{\sum_{i=1}^{N} Y_i^j} \quad (12)
\]

\[
F_{1-\text{macro}} = \frac{1}{n} \sum_{j=1}^{n} \frac{2 \sum_{i=1}^{N} Y_i^j Z_i^j}{\sum_{i=1}^{N} Y_i^j + \sum_{i=1}^{N} Z_i^j} \quad (13)
\]

Among the seven metrics, accuracy measures how many of the testing samples are recognized correctly in terms of both negative and positive samples. F1 score measures the weighted recall and precision, and is a more comprehensive metric than recall and precision. When comparing micro averaging F1-measure and macro averaging F1-measure, it is important to note that by definition, the macro averaged F1 would be more affected by the performance of classes with fewer examples, and micro averaged F1 would be more affected by the performance of the classes with more examples [38].

### C. Experimental Results of Emotion Tagging for Multimedia

For the Music Emotion database and the NVIE video database, Wang et al. [48] performed multiple emotion tagging using BN. Therefore, for these two databases, we directly compare our results to theirs et al. [48]. For the Memorability database and the FilmStim database, we conducted three comparison experiments.

Table II summarizes the experimental results of emotion tagging for multimedia, which also includes that of MLkNN [66] and BPMLL [65].

From Table II, we can observe the following.

First, our method significantly outperforms the feature-driven method on the NVIE video database, since the TRBM’s performance of the six of all the seven metrics is much better than that of the feature-driven method. On the other three databases, the performance of our method is much better than the feature-driven method for all the seven metrics. Therefore, for all four databases, our method outperforms the feature-driven method in terms of both label-based and sample-based metrics in almost every instance.

Second, compared to the model-based method, our method outperforms Wang et al.’s et al. [48] in terms of all of the five metrics in the Music Emotion database. It can be concluded that, on the Music Emotion database, TRBM captures much more effective relations among the multi-labels than the model-based method. Wang et al. use BN to capture the relations among multi-labels, and BN can capture pairwise dependency among labels, while TRBM can capture higher-order relations. On the NVIE video database, our method outperforms theirs on three of the five metrics, including sample-based recall, label-based micro-recall and label-based micro-F1 score. Therefore, we can conclude that the relations captured by TRBM on NVIE video database can help recognize more positive samples, which is the category with much fewer samples than the negative category. Since the experimental results of our method on the NVIE
TABLE II
EXPERIMENTAL RESULTS ON FOUR DATABASES

<table>
<thead>
<tr>
<th>Database</th>
<th>Method</th>
<th>Sample-based Accuracy</th>
<th>Sample-based Recall</th>
<th>Sample-based F1</th>
<th>Label-based Accuracy</th>
<th>Label-based Recall</th>
<th>Label-based F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memorability</td>
<td>BN</td>
<td>0.3462</td>
<td>0.4796</td>
<td>0.4525</td>
<td>0.5441</td>
<td>0.5729</td>
<td>0.5144</td>
</tr>
<tr>
<td></td>
<td>BPMLL</td>
<td>0.3503</td>
<td>0.5480</td>
<td>0.4794</td>
<td>0.5752</td>
<td>0.5398</td>
<td>0.5232</td>
</tr>
<tr>
<td></td>
<td>MLKNN</td>
<td>0.3213</td>
<td>0.3547</td>
<td>0.4426</td>
<td>0.4715</td>
<td>0.5114</td>
<td>0.4113</td>
</tr>
<tr>
<td></td>
<td>TRBM</td>
<td>0.3978</td>
<td>0.6699</td>
<td>0.5018</td>
<td>0.7326</td>
<td>0.5846</td>
<td>0.7483</td>
</tr>
<tr>
<td></td>
<td>feature-driven</td>
<td>0.5138</td>
<td>0.5981</td>
<td>0.5931</td>
<td>0.5957</td>
<td>0.6496</td>
<td>0.5830</td>
</tr>
<tr>
<td>Music Emotion</td>
<td>Wang et al. [50]</td>
<td>0.5520</td>
<td>0.6844</td>
<td>0.6293</td>
<td>0.6808</td>
<td>0.6600</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BPMLL</td>
<td>0.2602</td>
<td>0.3791</td>
<td>0.3593</td>
<td>0.3529</td>
<td>0.3685</td>
<td>0.2798</td>
</tr>
<tr>
<td></td>
<td>MLKNN</td>
<td>0.3648</td>
<td>0.4087</td>
<td>0.4409</td>
<td>0.3980</td>
<td>0.4875</td>
<td>0.3659</td>
</tr>
<tr>
<td></td>
<td>TRBM</td>
<td>0.5537</td>
<td>0.7476</td>
<td>0.6480</td>
<td>0.7536</td>
<td>0.6705</td>
<td>0.7438</td>
</tr>
<tr>
<td></td>
<td>feature-driven</td>
<td>0.3353</td>
<td>0.4076</td>
<td>0.4134</td>
<td>0.4148</td>
<td>0.4756</td>
<td>0.3596</td>
</tr>
<tr>
<td>NVIE video</td>
<td>Wang et al. [50]</td>
<td>0.4272</td>
<td>0.4752</td>
<td>0.4876</td>
<td>0.4375</td>
<td>0.4873</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BPMLL</td>
<td>0.3039</td>
<td>0.5907</td>
<td>0.4149</td>
<td>0.5682</td>
<td>0.4608</td>
<td>0.5155</td>
</tr>
<tr>
<td></td>
<td>MLKNN</td>
<td>0.4399</td>
<td>0.5271</td>
<td>0.5221</td>
<td>0.5514</td>
<td>0.5488</td>
<td>0.4484</td>
</tr>
<tr>
<td></td>
<td>TRBM</td>
<td>0.3179</td>
<td>0.6924</td>
<td>0.4431</td>
<td>0.7102</td>
<td>0.4901</td>
<td>0.7548</td>
</tr>
<tr>
<td>FilmStim</td>
<td>BN</td>
<td>0.2857</td>
<td>0.3451</td>
<td>0.3702</td>
<td>0.3684</td>
<td>0.4286</td>
<td>0.3022</td>
</tr>
<tr>
<td></td>
<td>BPMLL</td>
<td>0.2889</td>
<td>0.3583</td>
<td>0.4128</td>
<td>0.3918</td>
<td>0.4573</td>
<td>0.3113</td>
</tr>
<tr>
<td></td>
<td>MLKNN</td>
<td>0.3509</td>
<td>0.4661</td>
<td>0.4679</td>
<td>0.5000</td>
<td>0.5568</td>
<td>0.4507</td>
</tr>
<tr>
<td></td>
<td>TRBM</td>
<td>0.3818</td>
<td>0.7560</td>
<td>0.5560</td>
<td>0.7719</td>
<td>0.5546</td>
<td>0.7866</td>
</tr>
</tbody>
</table>

For the video database, our method captures more effective relations among multi-labels than the Bayesian network and enhances multi-label recognition on the NVIE video database and the Music Emotion database.

For the Memorability database and the FilmStim database, we compute all seven metrics for our TRBM and their BN method. From Table II, we find that on both of the databases, our method outperforms theirs, since for all seven metrics, the values of our method are higher than those of their method, indicating the better performance. We can conclude that our method captures more effective relations among multi-labels than the Bayesian network for enhancing multi-label recognition on the NVIE video database and the Music Emotion database.

Compared with BN, the proposed TRBM can capture high-order dependencies among multiple emotion labels, and hence improve the performance of emotion tagging. However, TRBM contains many more parameters than BN and it hence requires many more training data. Since BN requires both structure and parameters learning while TRBN only requires parameter learning yet with many more parameters to learn, the learning time of both models are equivalent. As for testing time, TRBM model takes longer time than the BN model due to the sparseness with the BN model.

TRBM significantly outperforms both BPMLL and MLKNN on the Memorability database and the Music Emotion database, since TRBM performs better than the other two methods for all seven metrics, and TRBM outperforms both BPMLL and MLKNN on four of the seven metrics of the NVIE Video database and the FilmStim database. Among the four databases, the Memorability database contains 2222 samples and the Music Emotion database contains 593 samples, while the NVIE Video database contains 72 samples and the FilmStim database contains 64 samples. Therefore, we can conclude that when enough samples are used, TRBM outperforms both BPMLL and MLKNN. In addition, for the NVIE Video database and the FilmStim database, TRBM is not the best among the methods in three of the seven metrics, including sample-based accuracy, sample-based F1 score and label-based micro-F1 score, but it is also not the worst. For the three metrics, BPMLL performs best among the three methods on the FilmStim database, while MLKNN performs best among the three methods on the NVIE Video database. Therefore, neither BPMLL nor MLKNN performs stably. So, on the whole, TRBM more consistently outperforms both BPMLL and MLKNN.

From the improved performance of BN and TRBM over the feature-driven method on the recall and F1 score metrics, we can conclude that compared to the feature-driven method, which independently recognizes each label and ignores the relation among multiple labels, the relations among the labels can facilitate the recognition of positive samples. These relations alleviate the database bias, and make the recognition much more effective, since both F1 score and recall improve.

These four databases include video samples, image samples, and music samples, as well as anywhere from 6 to 12 multi-labels. Our method works well for all of the sample types and multi-label numbers, which verifies that our method can be used for general multi-label recognition problems, and is not limited in multi-emotion tagging by sample type or multi-label number.

D. Semantic Relationship Analysis

In addition to evaluating the multi-label recognition performance of the proposed model, we show the captured semantic multi-label relationships. As discussed in Section III-C, each latent unit captures a specific multi-label presence or absence pattern that is implied by the parameters $W_{ij}$. Large $W$ indicates high probability of occurrence and small $W$ indicates high probability of absence.

Several patterns captured by the latent units of TRBM are shown in Figs. 4–7. In the four figures, the X-axis represents the multi-labels, and the Y-axis represents the weights $W_{ij}$. In order to show the figures clearly, the sigmoid function of $-0.5$ is used as the Y-axis. Unlike BN, which can only encode pairwise multi-label dependencies, the proposed model captures higher-order presence or absence patterns that involve all the multi-labels. Specifically, Fig. 4 shows that on the Music Emotion database, there exists one latent unit representing the presence of sadness, anger, relaxation, and quietness, with absence of amazement and happiness. From Fig. 5, we observe that on the NVIE video database, there is one latent unit that encodes a pattern in which a...
Fig. 4. Semantic multi-label relationship patterns captured by two latent units of TRBM on the Music Emotion database. (a) Pattern 1. (b) Pattern 2.

Fig. 5. Semantic multi-label relationship pattern captured by a latent unit of TRBM on the NVIE Video database.

Fig. 6. Semantic multi-label relationship patterns captured by two latent units of TRBM on the Memorability database. (a) Pattern 1. (b) Pattern 2.

Fig. 7. Semantic multi-label relationship pattern captured by a latent unit of TRBM on the FilmStim database.

video sample can make a person feel high levels of surprise and fear, some anger, sadness, and disgust, and neither happiness nor valance at all. From Fig. 6 shows one latent unit pattern on the Memorability database, which shows a higher degree of happiness than the other emotions, accompanied by a little bit of interest, arousal, strangeness, and fun, but neither attraction nor mystery. Fig. 7, we can see that on the FilmStim database, there exists one latent unit capturing one pattern in which both tenderness and joy are present and neither fear nor anger exists.

The pairwise relations captured by BN are shown in Fig. 8. From Fig. 8(a), we can find that on the music emotion database, there exist co-occur relation between sadness and quietness, and mutual exclusion relations between anger and happiness, relaxation and amazement, amazement and sadness, and amazement and quietness. For the NVIE video database, fear and disgust, surprise and disgust, disgust and anger, anger and sadness, and happiness and valance are co-occurrence, while fear and happiness, and sadness and happiness are mutual exclusive as shown in Fig. 8(b). For the Memorability database, the learned BN can capture four mutual exclusion relations and
Fig. 8. Learned BN structure on the four database. “E” and “M” preceding the emotion label separately represents true label and its measurement. (a) Music Emotion database. (b) NVIE video database. (c) Memorability database. (d) FilmStim database.

17 co-occur relations, corresponding the four dash links and 17 solid links respectively in Fig. 8(c). Fig. 8(d) indicates mutual exclusive relations between disgust and tenderness, tenderness and sadness, and fear and joy, as well as co-occurrence between disgust and anger, disgust and fear, and anger and sadness.

Comparing the performance of the TRBM model to the BN models, shown in Fig. 8, shows that for all four databases, all of the pairwise relationships captured by BN can be captured by our TRBM model. In practicality, for the Music Emotion database, all of the co-occurrence and mutually exclusive pairs of Fig. 8(a) can be found from Fig. 4. Both of the patterns captured by our TRBM model shown in Fig. 4 capture the co-occurrence pair of (quietness, sadness) shown in Fig. 8(a), as well as the mutually exclusive pairs, including (amazement, quietness), (amazement, sadness) and (relaxation, amazement). The other pairwise relationships are captured by either Fig. 4(a) [mutually exclusive pairs of (happiness, sadness)] or Fig. 4(b) [mutually exclusive pairs of (amazement, happiness), (relaxation, anger) and (anger, quietness)]. For the NVIE Video database, all of the co-occurrence and mutually exclusive pairs captured by Fig. 8(b) are also captured by the latent unit, as shown in Fig. 5, including the co-occurrence pairs of (fear, disgust), (surprise, disgust), (anger, disgust), (happiness, valence), and (sadness, disgust), as well as the mutually exclusive pairs of (fear, happiness) and (sadness, happiness). Comparing Fig. 8(c) and Fig. 6, similar conclusion can be obtained that the two patterns of TRBM capture all the pairwise relations showed in Fig. 8(c). A comparison between Fig. 8(d) and Fig. 7, shows that on the FilmStim database, one pattern captured by our TRBM model includes all the relations pairs captured by BN.

Our TRBM model can capture further dependencies among emotions in addition to these pairs. For example, on the Music Emotion database, our TRBM model also captures the mutually exclusive pairs of (happiness, anger), (valence, fear) and so on. On the NVIE video database, our TRBM model also captures much more other relations among the multi-labels, such as the mutually exclusive pair of (happiness, quietness).

V. CONCLUSION

Most current emotion tagging research tags the multimedia data with a single emotion, ignoring the dependencies among emotions. In this work, we propose a unified probabilistic framework for multiple emotion media tagging. First, the measurements are obtained using a traditional multi-label classification method. Then, TRBM is used to automatically model the higher-order dependencies among emotion tags. The experimental results on four multi-label databases show that our approach can effectively capture the higher-order relations among multi-labels. We compare our experimental results with those of the traditional feature-driven method, the method with BN capturing the pairwise dependencies among multi-labels, and other common multi-label recognition methods, and show that our method performs the best among these methods. Therefore, it further verifies that the relations modeled by our approach are more flexible than the pairwise relations captured by Bayesian network.

Four databases are adopted in this study: the Music Emotion database [41], the NVIE Video database [48], the Memorability database of Isola et al. [21], and the FilmStim database [35]. The sample number of these four databases ranges from 64 to 2222, and the number of multi-labels ranges from 6 to 12. The experimental results of our method on all four of the databases are much better than those of the feature-driven method, which
verifies that our method works efficiently under a variety of conditions.

REFERENCES


Jun Wang received the B.S. in computer science from the China University of Geosciences, Wuhan, China, in 2012, and the M.S. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2015.

Her research interest is affective computing.

Jun Wang received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree in electrical, computer, and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2015.

His research interests include machine learning and probabilistic graphical models and the applications in various areas, such as computer vision and brain computer interface.

Ziheng Wang received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree in electrical, computer, and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2015.

His research interests include machine learning and probabilistic graphical models and the applications in various areas, such as computer vision and brain computer interface.

Shangfei Wang (S’01–A’02–M’03–SM’15) received the B.S. degree in electronic engineering from Anhui University, Hefei, China, in 1996, the M.S. degree in circuits and systems from the University of Science and Technology of China (USTC), Hefei, China, in 1999, and the Ph.D. degree in signal and information processing from USTC in 2002.

From 2004 to 2005, she was a Postdoctoral Research Fellow with Kyushu University, Fukuoka, Japan. From 2011 to 2012, she was a Visiting Scholar with the Rensselaer Polytechnic Institute, Troy, NY, USA. She is currently an Associate Professor with the School of Computer Science and Technology, USTC. She has authored or coauthored over 70 publications. Her research interests cover computation intelligence, affective computing, and probabilistic graphical models.

Dr. Wang is a Member of the ACM.