Multimed Tools Appl (2015) 74:1863–1883
DOI 10.1007/s11042-013-1722-3

Multiple emotional tagging of multimedia data by exploiting dependencies among emotions

Shangfei Wang · Zhaoyu Wang · Qiang Ji

Published online: 12 October 2013

Abstract Digital multimedia may elicit a mixture of human emotions. Most current emotional tagging research typically tags the multimedia data with a single emotion, ignoring the phenomenon of multi-emotion coexistence. To address this problem, we propose a novel multi-emotion tagging approach by explicitly modeling the dependencies among emotions. First, several audio or visual features are extracted from the multimedia data. Second, four traditional multi-label learning methods: Binary Relevance, Random k label sets, Binary Relevance k Nearest Neighbours and Multi-Label k Nearest Neighbours, are used as the classifiers to obtain the measurements of emotional tags. Then, a Bayesian network is automatically constructed to capture the relationships among emotional tags. Finally, the Bayesian network is used to infer the data's multi-emotion tags by combining the measurements obtained from those traditional methods with the dependencies among emotions. Experiments on two multi-label media data sets demonstrate the superiority of our approach to the existing methods.

Keywords Multiple emotional tagging · Multimedia · Bayesian network · Multi-label classification

S. Wang (✉) · Z. Wang
Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology,
University of Science and Technology of China,
Hefei, Anhui 230027, People's Republic of China
e-mail: sfwang@ustc.edu.cn

Z. Wang
e-mail: wazhy@mail.ustc.edu.cn

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

1 Introduction

Recent years have seen a rapid increase in the size of digital multimedia collections, such as music, videos and images. Because emotion is an important component in the human classification and retrieval of digital media, assigning emotional tags to multimedia data has been an active research area in recent decades [17, 20, 50, 51, 62].

Previous research on multimedia emotional tagging mainly recognizes the emotional tags from the multimedia content. They can be summarized into two groups according to the adopted emotional tags: discrete categories and continuous dimensions. The former annotates multimedia using discrete emotional categories, such as calmness, happiness and fear [18, 19, 23, 28, 30, 43, 48, 53–58, 63]. The latter maps multimedia to continuous emotional dimensions, such as valence and arousal [1, 7, 13, 27, 37, 49, 66]. The framework of present research is as follows: first several audio or visual features are extracted from multimedia, then classification methods (e.g. Support Vector Machines (SVM)) or regression methods (e.g. Support Vector Regression (SVR)) are used to infer the multimedia data's emotional tag as either an emotional category or an emotional value in terms of valence and arousal.

The assumption of most present research is that one medium only has one emotional tag or a point in emotional dimensional space. However, most multimedia data often induce a mixture of emotions of users [12]. For example, a shocking video may elicit both anger and sadness; a piece of music may be characterized by both dreamy and cheerful. Some emotions may appear together frequently, while others may not. For example, Fig. 1 is a beautiful scene image, which may induce a mixed emotion of relaxing, comfortable and happy, but it rarely induces disgust. Such phenomena of co-existent and mutual exclusive relationships among emotions should be considered in emotional tagging. One medium should be assigned to several emotional tags simultaneously. Thus, we formulate emotional tagging as a multi-label classification problem.

Presently, few researchers regard emotional tagging of multimedia as a multi-label classification problem, except for a small number of studies on emotion recognition from music data. Furthermore, present multi-label classification methods, which

Fig. 1 A beautiful scene image.The image may induce a mixture of emotions including relaxing, comfortable and happy

address label dependencies directly either ignore the label correlations or fix the relations as a pairwise or a subset of label combinations existing in the training data. They cannot effectively explore the co-existence and mutual exclusion relationships among emotional labels. In this paper, the dependencies among emotional tags are explored directly by a Bayesian Network (BN).

In this paper, a novel approach named MET (Multiple Emotional Tagging of multimedia data by exploiting emotion dependencies) is proposed. First, several commonly used multi-label classifiers are adopted to obtain the measurements of the emotional tags from the audio-visual content. Then a BN is automatically constructed to model the dependencies among emotional tags. After that, the constructed BN is employed to infer the true tags for a medium based on the measurements. We conduct experiments on a multiple emotion music data set and a multiple emotion video data set. Experimental results show that MET exploits the co-existence and mutual exclusion relationships among emotions successfully. Thus, our method can improve the performance of traditional multi-label classifiers.

2 Related work

2.1 Emotional tagging of multimedia

Emotional tagging of videos, images and music pieces have attracted more and more attention in recent years [17, 20, 50, 51, 62]. There are two kinds of emotional tags, the expected emotion and the actual emotion [13]. The expected emotion is contained in a multimedia data and intended to be communicated toward users from multimedia program directors. It is likely to be elicited from majority of the users while consuming that multimedia. It can be considered as a common emotion. In contrast, the actual emotion is the affective response of a particular user to multimedia data. It is context-dependent and subjective, and it may vary from one individual to another. It can be considered as an individualized emotion. Most current research focus on the expected emotion, which is also the focus of this paper. Among the three kinds of media, the study of emotion recognition from music pieces has been carried out most profoundly, since almost every music piece is created to convey emotion. Emotional tagging of images was first studied in Japan in the 1990s [38]. At that time, Japanese word Kansei was used instead of emotional or affective. Emotional tagging of videos originated from the beginning of this century by Chitra Dorai, who proposed Computational Media Aesthetics (CMA) [6].

Although music pieces, images and videos are in different modalities, the research of emotional tagging of these three media obeys a similar framework. First, several discrete emotional categories or continuous emotional dimensions are adopted to express emotions. Second, audio or visual features are extracted from multimedia. After that, classification methods or regression methods are used to assign the media with an emotional tag or a point in emotional dimension space.

To express emotional categories, besides six basic emotions (i.e. happiness, sadness, surprise, fear, disgust and anger), adjectives and adjective pairs, such as pleasing, boring, and irritating, are often used. A famous categorical approach, Hevner's adjective checklist [14] is also adopted especially for music pieces. Some research tags the media into several discrete clusters using the clustering methods on

the arousal and valence spaces [24, 65]. To express continuous emotional dimensions, valence and arousal are often used for video and music tagging [1, 7, 13, 27, 37, 49, 66], while aesthetics [4] or attractiveness [2] is used for images.

Empirical research shows that the commonly used music features are timbre, rhythm, and harmony, which are associated with emotion perception of music [62]. For images, color, shape, and texture are extracted [17]. The video features contain both visual and audio features. The commonly used audio features include Mel-frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and spectral flux etc [21]. Shot duration, visual excitement, lighting key and color energy are widely used visual features [48].

Several machine learning algorithms have been applied to learn the relationships between features and discrete emotional labels, such as Gaussian mixture models [23], DBN [1], SVM [3], Neural Network [8] and conditional random fields [59] etc. After training, the automatic model can be applied to recognize the emotion of the media. For the continuous emotional dimension modeling, support vector regression [25], multiple linear regression [60], or AdaBoost.RT [60] are used to learn regression models to predict the valence and arousal values of the media. Most existing works train two regressions for valence and arousal independently [26, 61]. A comprehensive overview on emotional tagging of music pieces, images and videos can be found in [17, 20, 50, 51, 62].

To the best of our knowledge, present research of emotional tagging of images and videos assumes that there is only one emotional tag or a point in emotional dimensional space for an image or a video. However, it is very hard to find one video or image that can only induce a high level of a single emotional category without the presence of other emotions either in day-to-day living or inside the laboratory. Take videos for examples, Gross et al. [12] developed a set of films to elicit eight emotion states. Based on their study, when the users watch amusement videos, they always feel amused, happy and surprised simultaneously. The videos that induce anger may also induce some degree of disgust, sadness, fear and surprise. The videos that induce disgust may also induce fear and surprise to some extent. However, the videos that induce anger and disgust may not induce high level of happiness. Those phenomena of co-existence and mutual exclusion for emotional categories are also revealed in [32]. Till now, there is little research considering multi-emotion tagging of images or videos [52]. Thus, in this paper, we treat emotional image and video tagging as a multi-label classification problem.

For emotion recognition from music, there exists a small number of studies considering assigning multiple emotion labels to a music piece [29, 34, 35, 45]. However, the methods used in these studies, such as multi-label SVM [22] and Multi-Label k Nearest Neighbours (MLkNN) [45], either ignore the label correlations or fix the relations as a pairwise or a subset labels combinations in the training data. They cannot effectively exploit the coexistent and mutual-exclusive relations among emotions. Thus, we propose a BN to systematically capture the dependencies among emotional tags, which extend the relations modeled by current multi-label classifiers.

Besides viewing the emotions in terms of categories, much research assumes emotions have a systematic, coherent, and meaningful structure that can be mapped to affective dimensions [33, 39, 40]. Among those dimensions, arousal-valence (pleasure or activation) are always used. There are certain relationships between the emotional categories and dimensions. For example, in Russell's affective model, happiness

always belongs to the first quadrant of the model, and anger and fear belong to the fourth quadrant clearly. By considering the relationships between arousal-valence and emotion categories, one certain emotional category may be assigned to one of the four quadrants. In each quadrant, the emotional categories may be distinguished further using the relationships among emotional categories and dimensions. Although the continuous emotional dimensions provide more information for the media, they are difficult to be labeled and evaluated. Thus, we discretize emotional dimensions into two categories, positive and negative valence, or high and low arousal. Then, the relationships among emotional categories as well as the relationships between emotional categories and emotional dimensions are both taken into consideration in this paper.

2.2 Multi-label classifications

Multi-label classification is the classification problem where one sample can be assigned to more than one target label simultaneously. Multi-label classification methods can be categorized into two different groups: problem transformation methods and algorithm adaptation methods. The former includes Binary Relevance (BR) [47], Label Power (LP) [47], Random k labelsets (RAkEL) [46], etc. They transform the multi-label classification task into one or more single-label classification tasks and then any traditional classification algorithms can be used. The latter consists of Binary Relevance k Nearest Neighbours(BRkNN) [42], Multi-Label k Nearest Neighbours(MLkNN) [64], AdaBoost.MH [36] etc. They extend specific learning algorithms to handle multi-label data directly. A comprehensive overview of current research in multi-label classification can be found in [41].

Due to the large number of possible label sets, multi-label classification is rather challenging. Successfully exploring the coexistent and mutual exclusive relations inherent in multiple labels is the key to facilitate the learning process. Considering dependencies among labels, most present multi-label learning strategies can be categorized into three groups: methods ignoring label correlations, methods considering label correlations directly, and methods considering label correlations indirectly. The first group (i.e. BR) decomposes multi-label problem into multiple independent binary classification problems (one per category). Without considering the correlations among labels, the generalization ability of such method may be weak. The second group addresses the pairwise relations between labels(such as Calibrated Label Ranking (CLR) [9]), or the fixed label combinations present in training data (such as LP), or a random subset of the combinations(such as RAkEL)). However, the relations among labels may be beyond pairwise, and cannot be expressed by a fixed subset of labels existing in training data. Thus, the second group may not capture the label relations effectively. The third group considers label dependencies with the help of features or hypothesis. Godbole and Sarawagi [11] stacked the outputs of BR along with the full original feature space into a separate meta classifier, creating a two-stage classification process. Read et al. [15] proposed the classifier chain model to link n classifier into a chain. The feature space of each classifier in the chain is extended with the label associations of all previous classifiers. Ghamrawi and McCallum [10] adopted conditional random field to capture the impact of an individual feature on the co-occurrence probability of a pair of labels. Sun et al. [44]

proposed to construct a hyperedge for each label, and include all instances annotated with a common label into one hyperedge, thus capturing their joint similarity. Zhangs [67] proposed a Bayesian Network to model the dependencies among label errors, and then a binary classifier was constructed for each label combining the features and the parental labels, which were regarded as additional features. Huang et al. [16] modeled 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 hypothesis, these methods can model the flexible dependencies among labels to some extent, but their computation costs are usually much higher compared with the second group, which models the dependencies among labels directly.

Among the above, Zhang et al.'s work [67] is the most similar one to ours. Zhang et al. proposed to use a BN structure to encode the conditional dependencies of labels as well as the feature set: $P(\lambda_1, \lambda_2, ..., \lambda_n|x)$, where x is the features and $(\lambda_1, \lambda_2, ..., \lambda_n)$ are the multiple target labels, n is the number of labels. Since Zhang et al. thought directly modeling $P(\lambda_1, \lambda_2, ..., \lambda_n|x)$ by Bayesian approach was intractable, they adopted an approximate method to model the dependencies among label errors, which was independent of features x. Based on the learned BN structure of errors, a binary classifier was constructed for each label λ_i combining the features x and the parental labels $pa(\lambda_i)$, which were regarded as additional features.

Unlike Zhang et al.'s method, we propose a BN to systematically capture the dependencies among different labels, $P(\lambda_1, ..., \lambda_n)$, directly, without the help of features or hypothesis. The nodes of the BN represent the labels. The links and their parameters capture the probabilistic relations among labels. The label relationships encoded in a BN are more flexible than the pairwise or fixed subset of relationships used by the existing direct methods. The computation cost of our method is lower than that of indirect methods. Probabilistic reasoning model is used to infer the multiple labels with the largest probability in this paper.

Compared to related works, our contributions are as follows:

1. We propose a framework of multi-label multimedia emotional tagging, applicable to not only emotional tagging of music pieces, but also emotional tagging of images and videos. We are the first to formulate emotional tagging of images and videos as a multi-labeling problem.
2. We propose a novel method to automatically capture the dependencies among emotions directly with a BN and combine the captured emotion dependencies with their measurements to achieve accurate multi-emotion tagging of multimedia data.

3 Multiple emotional tagging methods

The framework of our approach is shown in Fig. 2, consisting of three modules: feature extraction, measurement extraction and multi-emotion relationship modeling by BN. The training phase of our approach includes training SVM and KNN in the traditional multi-label classification methods for measurement acquisition and training the BN to capture the semantic relationships among emotional tags. For measurement acquisition, we employ audio and visual features to represent the media, and then classify using traditional multi-label algorithms. Given the

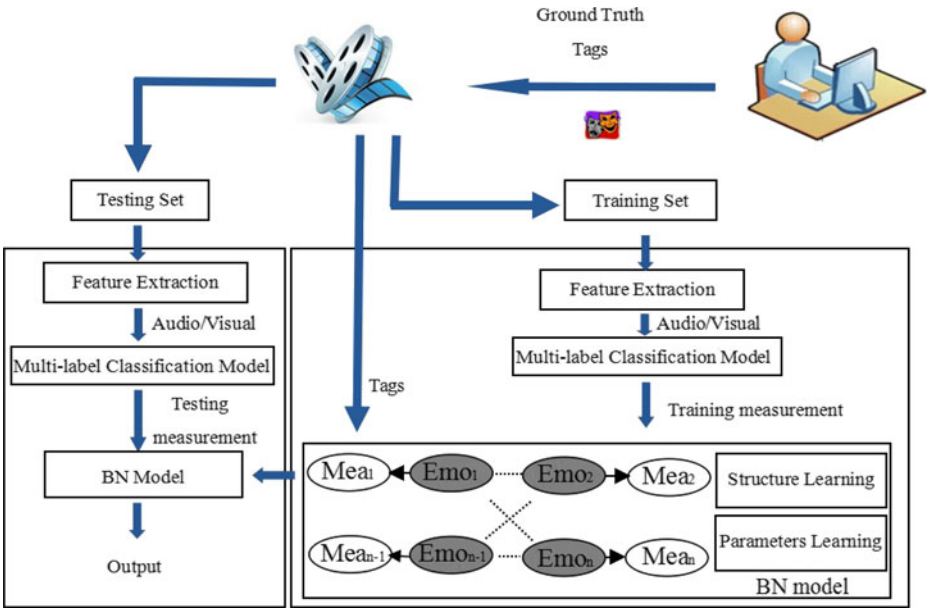

Fig. 2 The framework of our proposed emotional tagging approach

measurements, we infer the emotional tags of media through a probabilistic inference with the BN model. The details are provided as follows.

3.1 Feature extraction

Here, we only focus on the music and video features. Due to copyrights, the music data set [45] does not provide the original music clips, but 8 rhythmic features and 64 timbre features. The rhythmic features are derived by extracting periodic changes from a beat histogram. The timbre features consists 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 following sections.

For our collected video data set, visual and audio features are extracted. For visual features, three features, named lighting key, color energy and visual excitement [48], are extracted from video clips. These features are powerful tools to establish the mood of a scene and they can affect the emotions of the viewer according to cinematography and psychology. For audio features, we do not use the same features as music, since the audio part of videos include not only background music, but also speech and other sounds. Thirty-one features which are widely used in video tagging field [21] are extracted, including average energy, average energy intensity, spectrum flux, Zero Crossing Rate (ZCR), standard deviation of ZCR, 12 Mel-frequency Cepstral Coefficients (MFCCs), log energy of MFCC, 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 signals. These visual and

audio features are complementary for emotional tagging of videos. The details of features can be found in [21, 48].

3.2 Measurement acquisition

Four commonly used multi-label classification methods are adopted to obtain the measurements of the emotional tags. They are BR, RAkEL, BRkNN and MLkNN. The first two belong to problem transformation methods, and the last two belong to algorithm adaptation method. Below, we will briefly introduce the four methods.

Let $D = \{(x_i, y_i)\}_{i=1}^{m}$ represent the training data, where $x_i \in R^d$ is the feature, $y_i \in \{\lambda_j\}_{j=1}^{n}$ is the multiple target labels, n is the number of labels, and m is the number of training samples.

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_i for one label λ_i. Then, any traditional classification algorithm can be used to obtain the classifiers h_i using D_i. For a new instance, each classifier h_i outputs a binary label $Z_i = h_i(X_{fea})$. Then, the combination of the labels predicted by n classifiers $(\bigcup_{i=1}^{n} Z_i)$ is adopted as the final output. BR assumes the labels are independent, ignoring the correlations among those labels.

Another commonly used problem transformation method is LP, which considers each distinct label combination existing in the training set as a different class of a single-label classification task. Any traditional classification algorithm can be used to obtain the single-label classifier. A possible drawback of LP method is that some classes are associated with very few training samples which makes the learning difficult. To deal with the problem, RAkEL is proposed by breaking the initial set of labels into l random subsets, each subset has k labels, and then employing LP to train l corresponding classifiers. For a new instance, its labels are the combination of all the LP classifiers, which calculates the mean of these predictions for each label and outputs a final positive decision. RAkEL considers the randomly selected combinations of labels, but it does not capture the probabilistic relations among labels, and it cannot represent their coexistent and mutual exclusive relationships.

BRkNN is an algorithm adaptation method. It is conceptually equivalent to use Binary Relevance followed by KNN. BRkNN extends the kNN algorithm so that independent predictions are made for each label, following a single search of the k nearest neighbors [42]. In this case, the complexity of BRkNN is $1/n$ of that using BR and KNN directly. BRkNN does not consider the dependencies among labels.

MLkNN is another algorithm adaptation method based on BR. It uses maximum a posteriori principle to find the final label set based on prior and posterior probabilities of each k nearest neighbor labels. MLkNN does not consider the dependencies among labels.

The outputs of the above four methods are binary vectors, indicating whether a music piece or a video has a certain emotional tag or not. The binary vector is used as the measurement for the BN model in the following step.

3.3 Emotional relationship modeling by Bayesian network

As traditional tagging methods treat each emotional category individually but do not consider their dependencies in the training set, some valuable information may

be lost. In order to model the semantic relationships among emotional categories, we utilize a BN model for inference of emotional tags. As a probabilistic graphical model, BN can effectively capture the dependencies among variables in data. In our work, each node of the BN is an emotional label, and the links and their conditional probabilities capture the probabilistic dependencies among emotions.

3.3.1 BN structure and parameters learning

The BN learning consists of structure learning and parameters learning respectively. The structure consists of the directed links among the nodes, while the parameters are the conditional probabilities of each node given its parents.

Given the data set of multiple target labels $DL = \{(y_i)\}_{i=1}^{m}$, where $y_i \in \{\lambda_j\}_{j=1}^{n}$, the structure learning is to find a structure G that maximize a score function. In this work, we employ the Bayesian Information Criterion (BIC) score function which is defined as follows:

$$Score(G) = \max_{\theta} log(p(DL|G, \theta)) - \frac{Dim_G}{2} logm \tag{1}$$

where the first term is the log-likelihood function of parameters θ with respect to data DL and structure G, representing the fitness of the network to the data; the second term is a penalty relating to the complexity of the network, and Dim_G is the number of independent parameters.

To learn the structure, we propose to employ our BN structure learning algorithm [5]. By exploiting the decomposition property of the BIC score function, this method allows learning an optimal BN structure efficiently and it guarantees to find the global optimum structure, independent of the initial structure. Furthermore, the algorithm provides an anytime valid solution, i.e., the algorithm can be stopped at any-time with a best current solution found so far and an upper bound to the global optimum. Representing state of the art method in BN structure learning, this method allows automatically capturing the relationships among emotions. Details of this algorithm can be found in [5].

After the BN structure is constructed, parameters can be learned from the training data. Learning the parameters in a BN means finding the most probable values $\hat{\theta}$ for θ that can best explain the training data. Here, let Y_i denotes a variable of BN and y_i represents a generic state of Y_i. Each variable has a state space Ω_{Y_i}, where $y_i \in \Omega_{Y_i}$. Let θ_{ijk} denote a probability parameter for BN, then,

$$\theta_{ijk} = P\left(y_i^k \mid pa^j(Y_i)\right) \tag{2}$$

where $i \in \{1, ..., n\}$, $j \in \{1, ..., r_i\}$ and $k \in \{1, ..., s_i\}$. Here n denotes the number of variables (nodes in the BN), r_i represents the number of the possible parent instantiations for variable Y_i, and s_i indicates the number of the state instantiations for Y_i. Hence, y_i^k denotes the k_th state of variable Y_i.

Based on the Markov condition, any node in a Bayesian network is conditionally independent of its non-descendants, given its parents. The joint probability distribution represented by BN can be denoted as: $P(y) = P(y_1, ..., y_n) = \prod_i P(y_i|pa(Y_i))$. In this work, the "fitness" of parameters θ and training data D is quantified by the log likelihood function $log(P(D|\theta))$, denoted as $L_D(\theta)$. Assuming the training data are independent, based on the conditional independence assumptions in BN, the log

likelihood function is shown in Eq. 3. where n_{ijk} indicates the number of elements in D containing both y_i^k and $pa^j(Y_i)$.

Because there is no label missing in training data in this work, Maximum Likelihood Estimation (MLE) method can be described as a constrained optimization problem, which is shown in Eq. 3.

$$MAX \quad L_D(\theta) = log\left(\prod_{i=1}^{n}\prod_{j=1}^{r_i}\prod_{k=1}^{s_i}\theta_{ijk}^{n_{ijk}}\right)$$
$$S.T \quad g_{ij}(\theta) = \sum_{k=1}^{s_i}\theta_{ijk} - 1 = 0 \tag{3}$$

where g_{ij} imposes the constraint that the parameters of each node sums to 1 over all the states of that node. Solving the above equations, we can get $\theta_{ijk} = \frac{n_{ijk}}{\sum_k n_{ijk}}$.

3.3.2 BN inference

During the BN inference, the posterior probability of categories can be estimated by combining the likelihood from measurement with the prior model. Let E_i and M_i, $i \in \{1, ...n\}$, denote the variable and the corresponding measurements obtained by machine learning methods respectively. Then,

$$P(E_1, ..., E_n \mid M_1, ..., M_n) = \prod_{i=1}^{n} P(M_i \mid E_i)\prod_{i=1}^{n} P(E_i \mid pa(E_i))) \tag{4}$$

The condition probability in the equation are learned from training set. In this work, the inferred tags are the emotion tag string $(E_1, ..., E_n)$ with the highest probability given $M_1, ..., M_n$. In practice, the belief propagation algorithm [31] is used to estimate the posterior probability of each category node efficiently.

4 Experiments and results

4.1 Experimental conditions

4.1.1 Data sets

Presently, there is only one available music data set with multiple emotional labels, and no multiple emotion image or video data set. Thus, in this work, we use two data sets: multiple emotion music data set [45] and multiple emotion video data set collected by us.

The music data set contains 593 songs categorized into one or more out of 6 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 s and the sampling rate of speech is 22.05 kHz. The distribution of samples is presented in Table 1. Detailed information about the data set can be found in [45]. Since the music data set does not provide the valence and arousal labels, we only model the category relations using BN.

Table 1 Sample distribution in music data set

Emotion	Amazed	Happy	Relaxing	Quiet	Sad	Angry
Number	173	166	264	148	168	189

For our constructed multi-label emotion video data set, we first obtain 72 videos which last 8166 s overall from internet as the stimulus. The lengths of the videos vary from half minute to five minutes. The sampling rate of the speech and video are 44 kHz and 30 fps. We assume there is no temporal change or transition of emotional experience within a single clip because of its short duration. These videos are grouped into several playlists and each playlist contained six video shots. To reduce the interaction between two consecutive target videos, a relaxing video approximately 1–2 min in length was shown between two target videos. More than fifty healthy students were recruited to participate in the experiment to watch each playlist.

After watching each video shot, subjects were asked to report their actual experienced emotions using emotional valence and arousal that range from −2 to 2, implying negative to highly positive valence and calm to exciting arousal, respectively. Subjects also rated the intensity of the six basic emotional categories for the video, which ranged from 0 to 4, where 0 indicates no particular feeling and 4 indicates a strong feeling. The average intensities of the self-reported data were used as the ground truth emotional tags for the videos.

After data collection, a threshold is set to transform the intensity of emotional tag to a binary tag, which represents 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 that for valence and arousal is 1. The sample distribution is presented in Table 2.

4.1.2 Evaluation metrics

For two problem transformation method, BR and RAKEL, the SVM with a linear kernel is used as the basic classifier. The measurements obtained using BR, RAkEL, BRkNN, MLkNN methods are regarded as the input of the BN to infer the final emotions. 10-fold cross-validation is adopted. For each fold, the four traditional multi-label learning methods and the BN share the same training set and testing set.

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 correctly or partly incorrectly. Thus, there are two kinds of commonly used metrics, example-based and label-based measures (see [41] for an explanation of both), evaluating the multi-label emotional tagging performance from the view of instances and labels respectively. We adopt both measures in this work. Let Y_i denotes the true labels for instance i, which is a binary vector, and Z_i is the predicted labels for instances i, m represents the number of the instances and n is the number of labels. The example-based measures: accuracy, precision, recall,

Table 2 Sample distribution in self-constructed video data set

Emotion	Happiness	Anger	Sadness	Fear	Disgust	Surprise	Valence
Number	28	12	17	34	29	27	29

F1-measure and subset accuracy are defined in Eqs. 5–9 [41], and the label-based measures: recall, precision and F1-measure, are defined in Eqs. 10–12 [41].

$$Accuracy = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{Y_i \bigcap Z_i}{Y_i \bigcup Z_i}\right| \tag{5}$$

$$Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{\left|Y_i \bigcap Z_i\right|}{\left|Z_i\right|} \tag{6}$$

$$Recall = \frac{1}{m}\sum_{i=1}^{m}\frac{\left|Y_i \bigcap Z_i\right|}{\left|Y_i\right|} \tag{7}$$

$$F_1 = \frac{1}{m}\sum_{i=1}^{m}\frac{2\left|Y_i \bigcap Z_i\right|}{\left|Y_i\right| + \left|Z_i\right|} \tag{8}$$

$$Subset\,Accuracy = \frac{1}{m}\sum_{i=1}^{m} I(Y_i = Z_i) \tag{9}$$

$$Precision,\ P_{micro} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} Y_i^j Z_i^j}{\sum_{j=1}^{n}\sum_{i=1}^{m} Z_i^j} \tag{10}$$

$$Recall,\ R_{micro} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m} Y_i^j Z_i^j}{\sum_{j=1}^{n}\sum_{i=1}^{m} Y_i^j} \tag{11}$$

$$F_{1-micro} = \frac{2\sum_{j=1}^{n}\sum_{i=1}^{m} Y_i^j Z_i^j}{\sum_{j=1}^{n}\sum_{i=1}^{m} Y_i^j + \sum_{j=1}^{n}\sum_{i=1}^{m} Z_i^j} \tag{12}$$

4.2 Experimental Results and Analyses of Emotional Tagging of Music

We quantify the co-occurrence among different emotional tags using a conditional probability of $P(B|A)$, where A is one emotional tag, and B is another emotional tag. $P(B|A)$ therefore measures the probability of emotional tag B happens, given emotion A happens. Table 3 shows the condition probabilities between different emotions for the music data set. From the table, each music piece can display multiple emotions. For instance, quiet is often accompanied by relaxing and sad with high probability. From the table, we can find clearly two kinds of relationships among

Table 3 Dependencies among emotional labels for the music dataset

P(B|A)		B					
		Amazed	Happy	Relaxing	Quiet	Sad	Angry
A	Amazed	1	0.3237	0.0751	0	0.0578	0.5318
	Happy	0.3373	1	0.5482	0.0422	0.006	0.0723
	Relaxing	0.0492	0.3447	1	0.3939	0.3598	0.0265
	Quiet	0	0.0473	0.7027	1	0.7095	0.0135
	Sad	0.0595	0.006	0.5655	0.625	1	0.119
	Angry	0.4868	0.0635	0.037	0.0106	0.1058	1

emotions, which are co-occurrent relationships and mutual exclusive relationships. For example, the probabilities of $P(angry|relaxing)$ and $P(amazed|relaxing)$ are 0.0265 and 0.0492, which indicate that relaxing rarely coexists with amazed and anger. $P(happy|angry)$ and $P(happy|sad)$ are 0.0635 and 0.006, which show happy rarely coexists with sad and angry. Quiet is always coexistent with sad as indicated by a high $P(sad|quiet)$ of 0.7095.

To systematically capture such relationships among emotions in the music data, we learnt a BN. Figure 3 shows the learned BN, where the shaded nodes are hidden nodes and they represent the true state we want to infer, and the unshaded nodes are the measurement nodes obtained from a traditional multi-labeling method. The links among the shaded nodes represent the dependencies among emotions. For example, the link from relaxing to angry and amazed demonstrates there are strong dependences between the two pairs. From the Table 3, we can see the probabilities of $P(angry|relaxing)$ and $P(amazed|relaxing)$ are 0.0265 and 0.0492, which indicates they are mutual exclusive relationship. Meanwhile, the link from quiet to sad shows the co-occurrent relationship because the probability of $P(sad|quiet)$ is 0.7095 in the Table 3. They demonstrate that the BN can effectively capture the mutual exclusive and coexistent relations among emotional labels. These kinds of relations among labels are beyond the scope of those captured by commonly used multi-label learning methods.

Using the BN, we can then infer the true emotion labels by instantiating the measurement nodes with the emotion estimates obtained from a traditional multi-labeling classification method. The inference results are summarized in Table 4.

Table 4 shows the performances of our approach and commonly used multi-label classifiers. From Table 4, we can obtain the following observations:

1. RAkEL performs the best among the four commonly used multi-label classifiers, which is consistent with the work [45] for music data set. The reason may be that RAkEL considers the relations of the randomly selected sub-combinations existing in the training label sets, where the other three methods ignore any

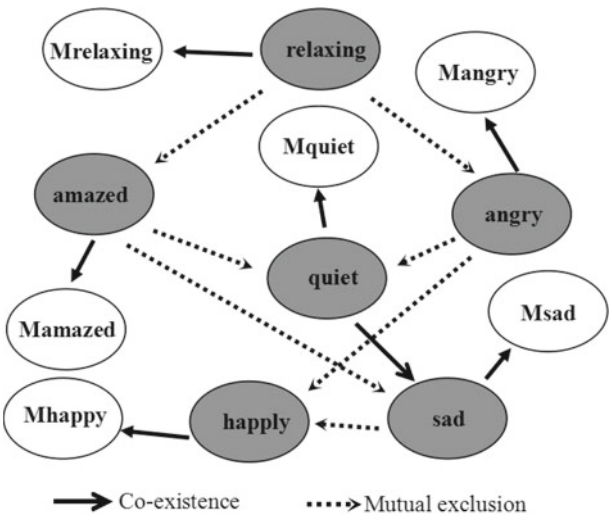

Fig. 3 The learned BN structure from music data set. The links among shaded nodes show the dependencies among emotions

Table 4 Results of comparison experiments of our model and commonly used multi-label classifiers in music data set

Method	Example based					Label based		
	Acc.	Pre.	Rec.	F1	SubAcc.	MicPre.	MicRec.	MicF1.
BR	0.5138	0.6501	0.5981	0.5931	0.2681	0.7143	0.5957	0.6496
BR + BN	0.5520↑	0.6234	0.6844↑	0.6293↑	0.3221↑	0.6352	0.6868↑	0.6600↑
Improvement	7.42 %	−4.11 %	14.43 %	6.11 %	20.13 %	−11.07 %	15.30 %	1.60 %
RAkEL	0.5719	0.7004	0.6703	0.6544	0.3238	0.7046	0.6652	0.6843
RAkEL + BN	0.5749↑	0.6667	0.7026↑	0.6580↑	0.3272↑	0.6618	0.7013↑	0.6810
Improvement	0.53 %	−4.82 %	4.82 %	0.54 %	1.04 %	−6.07 %	5.43 %	−0.49 %
BRkNN	0.5145	0.6650	0.5801	0.5903	0.2884	0.7346	0.5794	0.6478
BRkNN + BN	0.5541↑	0.6341	0.6844↑	0.6346↑	0.3204↑	0.6451	0.6823↑	0.6632↑
Improvement	7.70 %	−4.65 %	17.97 %	7.49 %	11.11 %	−12.18 %	17.76 %	2.37 %
MLkNN	0.5344	0.6782	0.6228	0.6177	0.2867	0.7100	0.6209	0.6625
MLkNN + BN	0.5562↑	0.6355	0.6925↑	0.6392↑	0.3137↑	0.6405	0.6931↑	0.6658↑
Improvement	4.08 %	−6.30 %	11.19 %	3.49 %	9.41 %	−9.79 %	11.63 %	0.50 %

"acc." refers to " accuracy","pre." refers to "precision","rec." refers to "recall", "subAcc." refers to "subsetAccuracy", "micPre." refers to "micro precision", "micRec." refers to "micro recall", "micF1." refers to "micro F1"

relations among labels. It proves the importance of label relations for multi-label classification.

2. Our approach outperforms the four commonly used multi-label classifiers, since both example based and label based measures of our approach are better than those of four commonly used multi-label classifiers in most cases. It demonstrates the effectiveness of our approach, since it can more effectively capture the dependencies among emotional labels. Furthermore, our method increases the example based accuracy, example based F1, and the label based F1 in most cases. It indicates that our method not only improves the recognition accuracy, but also makes the recognition results more balanced.
3. By using BN, the improvements for the four commonly used multi label classifiers are different. The improvements of most measures for BRkNN method are highest and those for RAkEL are lowest. RAkEL already consider the label relations to some extent, and other three methods do not. Thus the enhancement due to the relations modeled by BN, for RAkEL is less than others.

4.3 Experimental results and analyses of emotional tagging of videos

We performed a similar study for multi-emotion tagging of the video data. Table 5 shows the condition probabilities between emotional labels from video data set. From Table 5, we can also find emotional videos can induce multiple emotions. For instance, surprise is present with a high probability when happiness is present. Some degree of fear and surprise are present given disgust. Disgust, sadness, fear and surprise are always present when the video induces anger. Disgust and surprise may appear when fear is present. This is consistent with previous study results described in [12]. Two kinds of relationships among emotions, which are co-occurent relationship and mutual exclusive relationship are shown clearly in the table. For example, the probability of $P(valence|happiness)$ is 1 which means these two emotions occur together frequently and reflects the co-occurrent relationship. On the other hand,

Table 5 Dependencies among emotional labels for video data set

P(B|A)		B							
		Hap	Ang	Sad	Fea	Dis	Sur	Positive val.	High aro.
A	Hap	1	0	0	0.0714	0.1071	0.4643	1	1
	Ang	0	1	0.75	0.6667	0.75	0.5	0.0833	1
	Sad	0	0.5294	1	0.4118	0.5294	0.3529	0	1
	Fea	0.0588	0.2353	0.2059	1	0.6765	0.3824	0.0882	1
	Dis	0.1034	0.3103	0.3103	0.7931	1	0.5517	0.1034	1
	Sur	0.4815	0.2222	0.2222	0.4815	0.5926	1	0.4815	1
	Positive val.	0.9655	0.0345	0	0.1034	0.1034	0.4483	1	1
	High aro.	0.3889	0.1667	0.2301	0.4722	0.4028	0.375	0.4028	1

the probability of $P(happiness|\, fear)$ is 0.0588, which means there is few samples of admixture emotion of happiness and fear and indicates mutual exclusive relationship. Besides, from the table, the arousal of all videos is evaluated as high arousal, which indicates the videos we collected aroused the interest of the subject. Because there is no low arousal video, it can not use algorithms to get the measurement for arousal node in the experiment. Thus, in the following experiments, we can only obtain 7 measurements, that is, 6 basic emotional tags and valence.

Given the video data, we can then learn a BN to capture the relationships among the emotions. The learned BN is shown in Fig. 4. As discussed above, there is only high arousal video in the data set and the dependencies between arousal and other emotional tags are not very evident. So the arousal node is isolated in this structure. The links among the shaded nodes show the dependencies among emotions. For example, the link from fear to happiness demonstrates there is strong

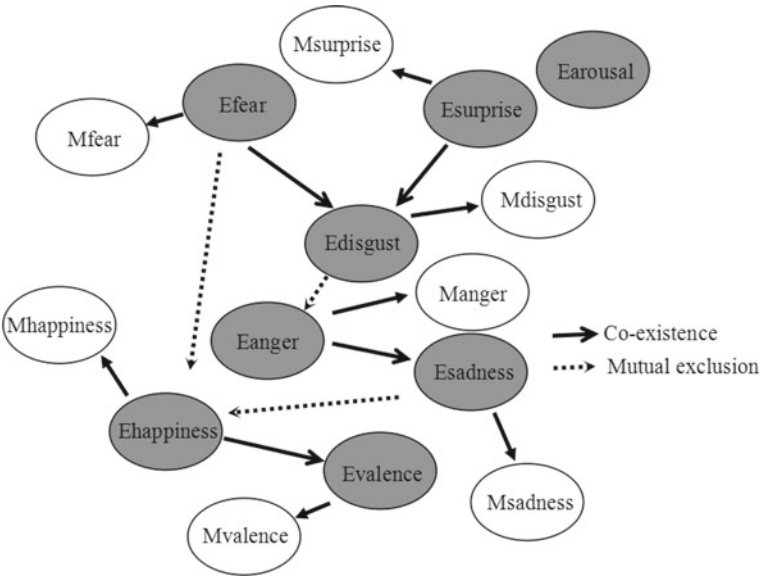

Fig. 4 The learned BN structure from video data set. The links among shaded nodes show the dependencies among emotions

Table 6 Results of comparison experiment of our model and commonly used multi-label classifiers in video data set

Method	Example based					Label based		
	Acc.	Pre.	Rec.	F1	SubAcc.	MicPre.	MicRec.	MicF1.
BR	0.4789	0.5828	0.6382	0.5757	0.1806	0.5860	0.6193	0.6022
BR + BN	0.5166↑	0.6315↑	0.6053	0.5906↑	0.2917↑	0.6279↑	0.6136	0.6207↑
Improvement	7.89 %	8.35 %	−5.15 %	2.59 %	61.54 %	7.15 %	−0.92 %	3.07 %
RAkEL	0.5693	0.6528	0.7183	0.6556	0.2917	0.6425	0.7045	0.6721
RAkEL + BN	0.5947↑	0.6900↑	0.7248↑	0.6745↑	0.3472↑	0.6578↑	0.6989	0.6777↑
Improvement	4.47 %	5.71 %	0.90 %	2.89 %	19.05 %	2.38 %	−0.81 %	0.83 %
BRkNN	0.3565	0.5208	0.4039	0.4225	0.1944	0.6476	0.3864	0.4840
BRkNN + BN	0.4333↑	0.5602↑	0.4889↑	0.4994↑	0.2639↑	0.5438	0.4943↑	0.5179↑
Improvement	21.56 %	7.56 %	21.03 %	18.22 %	35.71 %	−16.04 %	27.94 %	7.00 %
MLkNN	0.3353	0.4699	0.4076	0.4134	0.1389	0.5573	0.4148	0.4756
MLkNN + BN	0.4272↑	0.5486↑	0.4752↑	0.4876↑	0.2500↑	0.5500	0.4375↑	0.4873↑
Improvement	27.42 %	16.75 %	16.58 %	17.95 %	80.00 %	−1.30 %	5.48 %	2.48 %

"acc." refers to " accuracy","pre." refers to "precision","rec." refers to "recall", "subAcc." refers to "subsetAccuracy", "micPre." refers to "micro precision", "micRec." refers to "micro recall", "micF1." refers to "micro F1"

relationship between the pair. From the Table 5, we can see the probability of $P(happiness|\,fear)$ is 0.0588, which indicates it is mutual exclusive relationship. The link from happiness to valence shows the co-occurrent relationship because the probability of $P(valence|happiness)$ is 1 in the Table 5.

Table 6 shows the results of comparison experiments of our model and commonly used multi-label classifiers in video data set. From Table 6, we find that, among the four traditional methods, RAkEL again performs the best. Compared to those of traditional methods, MET shows significant performance improvement on most of the measures. The improvements of most measures for BRkNN and MLkNN method are the highest and those for RAkEL are the lowest. Similar to Section 4.2, these observations further prove that the importance of label relations for multi-label classification, and the effectiveness of our approach.

5 Conclusions

Most current emotional 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 four traditional multi-label classification methods which are BR, RAkEL, BRkNN and MLkNN. Second, BN is used to automatically model the dependencies among emotional tags. The experimental results on two multi-label data sets show that our approach can effectively capture the co-occurrence and mutual exclusive relations among emotions, and thus, our approach outperforms other methods. The relations modeled by our approach are more flexible than pairwise or fixed subset labels captured by current multi-label learning methods.

Two data sets are adopted in this study: the multiple emotion music data set [45] and the multiple emotion video data set collected by us. The size of these two data

sets is small, especially the multiple emotion video data set. The video data set is also constructed imbalanced, since it does not include any low arousal video. Large scale and balanced multi-label multimedia data set is a key requirement for research of multimedia emotional tagging. In the future, we will add low arousal videos and extend our multi-label emotion video data set.

Acknowledgements This paper is supported by the NSFC (61175037, 61228304), Special Innovation Project on Speech of Anhui Province (11010202192), Project from Anhui Science and Technology Agency(1106c0805008) and the Fundamental Research Funds for the Central Universities. We also acknowledge partial support from the US National Science Foundation under grant # 1205664.

References

1. Arifin S, Cheung PYK (2007) A novel probabilistic approach to modeling the pleasure-arousal-dominance content of the video based on "working memory". In: International Conference on Semantic Computing, ICSC 2007. IEEE, pp 147–154
2. Bhattacharya S, Sukthankar R, Shah M (2010) A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proceedings of the international conference on multimedia. ACM, pp 271–280
3. Bischoff K, Firan CS, Paiu R, Nejdl W, Laurier C, Sordo M (2009) Music mood and theme classification-a hybrid approach. In: Proceedings of the international conference on music information retrieval, pp 657–662
4. Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: Computer vision–ECCV 2006, pp 288–301
5. de Campos CP, Ji Q (2011) Efficient structure learning of Bayesian networks using constraints. J Mach Learn Res 12:663–689
6. Dorai C, Venkatesh S (2001) Computational media aesthetics: finding meaning beautiful. IEEE Multimedia 8(4):10–12
7. Eerola T, Lartillot O, Toiviainen P (2009) Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: Proceedings of the international conference on music information retrieval, pp 621–626
8. Feng Y, Zhuang Y, Pan Y (2003) Popular music retrieval by detecting mood. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. ACM, pp 375–376
9. Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153
10. Ghamrawi N, McCallum A (2005) Collective multi-label classification. In: Proceedings of the 14th ACM international Conference on Information and Knowledge Management, CIKM '05. ACM, New York, pp 195–200
11. Godbole S, Sarawagi S (2004) Discriminative methods for multi-labeled classification. In: Proceedings of the 8th Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 22–30
12. Gross JJ, Levenson RW (1995) Emotion elicitation using films. Cogn Emot 9(1):87–108
13. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimedia 7(1):143–154
14. Hevner K (1935) Expression in music: a discussion of experimental studies and theories. Psychol Rev 42(2):186
15. Holmes G, Read J, Pfahringer B, Frank E (2009) Classifier chains for multi-label classification. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: part II, ECML PKDD '09. Springer, Berlin, Heidelberg, pp 254–269
16. Huang S-J, Yu Y, Zhou Z-H (2012) Multi-label hypothesis reuse. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD '12. ACM, New York, pp 525–533
17. Joshi D, Datta R, Fedorovskaya E, Luong Q-T, Wang JZ, Li J, Luo J (2011) Aesthetics and emotions in images. IEEE Signal Process Mag 28(5):94–115
18. Kang HB (2003) Affective contents retrieval from video with relevance feedback. In: Sembok TMT, Zaman HB, Chen H, Urs SR, Myaeng S-H (eds) Digital libraries: technology and

management of indigenous knowledge for global access. Lecture Notes in Computer Science, vol. 2911. Springer Berlin Heidelberg, pp 243–252
19. Kim EY, Kim SJ, Koo HJ, Jeong K, Kim JI (2005) Emotion-based textile indexing using colors and texture. In: Wang L, Jin Y (eds) Fuzzy systems and knowledge discovery. Lecture Notes in Computer Science, vol. 3613. Springer Berlin Heidelberg, pp 1077–1080
20. Kim YE, Schmidt EM, Migneco R, Morton BG, Richardson P, Scott J, Speck JA, Turnbull D (2010) Music emotion recognition: a state of the art review. In: Proc. ISMIR. Citeseer, pp 255–266
21. Li D, Sethi IK, Dimitrova N, McGee T (2001) Classification of general audio data for content-based retrieval. Pattern Recogn Lett 22(5):533–544
22. Li T, Ogihara M (2003) Detecting emotion in music. In: Proceedings of the international symposium on music information retrieval. Washington, pp 239–240
23. Liu D, Lu L, Zhang H-J (2003) Automatic mood detection from acoustic music data. In: Proceedings of the international symposium on music information retrieval, pp 81–87
24. Liu C-C, Yang Y-H, Wu P-H, Chen HH (2006) Detecting and classifying emotion in popular music. In: Proceedings of the joint international conference on information sciences, pp 996–999
25. Luca C, Sergio B, Riccardo L (2013) Affective recommendation of movies based on selected connotative features. IEEE Trans Circ Syst Video Technol 23(4):636–647
26. MacDorman KF, Ough S, Hoa C-C (2007) Automatic emotion prediction of song excerpts: index construction, algorithm design, and empirical comparison. J New Music Res 36(4): 281–299
27. Marcelino R, Teixeira A, Yamasaki T, Aizawa K (2011) Determination of emotional content of video clips by low-level audiovisual features. Multimed Tools Appl 61(1):21–49
28. Moncrieff S, Dorai C, Venkatesh S (2001) Affect computing in film through sound energy dynamics. In: Proceedings of the ninth ACM international conference on multimedia. ACM, pp 525–527
29. Myint EEP, Pwint M (2010) An approach for mulit-label music mood classification. In: 2nd International Conference on Signal Processing Systems, (ICSPS) 2010, vol 1. IEEE, pp 290–294
30. Oliveira E, Martins P, Chambel T (2011) Ifelt: accessing movies through our emotions. In: Proceddings of the 9th international interactive conference on interactive television, EuroITV '11. ACM, pp 105–114
31. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann
32. Philippot P (1993) Inducing and assessing differentiated emotion-feeling states in the laboratory. Cogn Emot 7(2):171–193
33. Russell JA (1997) Reading emotion from and into faces: resurrecting a dimensional-contextual perspective. In: Russell JA, Fernández-Dols JM (eds) The psychology of facial expression. Studies in emotion and social interaction, 2nd series. Cambridge University Press, New York, NY, pp 295–320
34. Sanden C, Zhang JZ (2011) An empirical study of multi-label classifiers for music tag annotation. In: 12th international society for music information retrieval conference, (ISMIR 2011). Citeseer, pp 717–722
35. Santos AM, Canuto AMP, Neto AF (2011) A comparative analysis of classification methods to multi-label tasks in different application domains. Int J Comput Inform Syst Ind Manage Appl. 3:218–227
36. Schapire RE, Singer Y (2000) Boostexter: a boosting-based system for text categorization. Mach Learn 39(2):135–168
37. Schuller B, Johannes D, Gerhard R (2010) Determination of nonprototypical valence and arousal in popular music: features and performances. EURASIP Journal on Audio, Speech, and Music Processing 2010(735854):1–19. doi:10.1155/2010/735854
38. Shibata T, Kato T (1999) "Kansei" image retrieval system for street landscape-discrimination and graphical parameters based on correlation of two images. In: IEEE international conference on systems, man, and cybernetics, 1999. IEEE SMC'99 conference proceedings, vol 6. IEEE, pp 247–252
39. Smith CA (1989) Dimensions of appraisal and physiological response in emotion. J Personal Soc Psychol 56(3):339
40. Smith CA, Lazarus RS (1991) Emotion and adaptation. Oxford University Press, New York
41. Sorower MS (2010) A literature survey on algorithms for multi-label learning. Ph.D Qualifying Review Paper, Oregon State University

42. Spyromitros E, Tsoumakas G, Vlahavas I (2008) An empirical study of lazy multilabel classification algorithms. In: Darzentas J, Vouros GA, Vosinakis S, Arnellos A (eds) Artificial intelligence: theories, models and applications. Lecture Notes in Computer Science, vol. 5138. Springer Berlin Heidelberg, pp 401–406
43. Sun K, Yu J (2007) Video affective content representation and recognition using video affective tree and hidden markov models. In: Paiva ACR, Prada R, Picard RW (eds) Affective computing and intelligent interaction. Lecture Notes in Computer Science, vol. 4738. Springer Berlin Heidelberg, pp 594–605
44. Sun L, Ji S, Ye J (2008) Hypergraph spectral learning for multi-label classification. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD '08. ACM, New York, pp 668–676
45. Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I (2008) Multi-label classification of music into emotions. In: ISMIR 2008: proceedings of the 9th international conference of music information retrieval, pp 325–330
46. Tsoumakas G, Vlahavas I (2007) Random k-labelsets: An ensemble method for multilabel classification. In: Proceedings of the 18th European conference on machine learning, pp 406–417
47. Tsoumakas G, Katakis I, Vlahavas I (2010) Mining multi-label data. In: Maimon O, Rokach L (eds) Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer, New York, pp 667–685
48. Wang HL, Cheong L-F (2006) Affective understanding in film. IEEE Trans Circ Syst Video Technol 16(6):689–704
49. Wang CW, Cheng WH, Chen JC, Yang SS, Wu JL (2006) Film narrative exploration through the analysis of aesthetic elements. In: Cham TJ, Cai J, Dorai C, Rajan D, Chua TS, Chia LT (eds) Advances in multimedia modeling. Lecture Notes in Computer Science, vol. 4351. Springer Berlin Heidelberg, pp 606–615
50. Wang W, He Q (2008) A survey on emotional semantic image retrieval. In: 15th IEEE International Conference on Image Processing, ICIP 2008. IEEE, pp 117–120
51. Wang S, Wang X (2010) Emotional semantic detection from multimedia: a brief overview. In: Dai Y, Chakraborty B, Shi M (eds) Kansei engineering and soft computing: theory and practice 2010. IGI press, pp 126–146
52. Wang Z, Wang S, He M, Ji Q (2013) Emotional tagging of videos by exploring multi-emotion coexistence. In: IEEE international conference on automatic face & gesture recognition and workshops, (FG 2013). IEEE
53. Watanapa SC, Thipakorn B, Charoenkitarn N(2008) A sieving ANN for emotion-based movie clip classification. IEICE TRans Inf Syst 91(5):1562–1572
54. Wei CY, Dimitrova N, Chang SF (2004) Color-mood analysis of films based on syntactic and psychological models. In: IEEE International Conference on Multimedia and Expo, ICME'04, vol 2. IEEE, pp 831–834
55. Wei-ning W, Ying-lin Y, Sheng-ming J (2006) Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: IEEE international conference on Systems, Man and Cybernetics, SMC'06, vol 4. IEEE, pp 3534–3539
56. Winoto P, Tang TY (2010) The role of user mood in movie recommendations. Expert Syst Appl 37(8):6086–6092
57. Wu T-L, Jeng S-K (2008) Probabilistic estimation of a novel music emotion model. In: Satoh S, Nack F, Etoh M (eds) Advances in Multimedia Modeling. Lecture Notes in Computer Science, vol. 4903. Springer Berlin Heidelberg, pp 487–497
58. Xu M, He X, Jin JS, Peng Y, Xu C, Guo W (2011) Using scripts for affective content retrieval.In: Qiu G, Lam KM, Kiya H, Xue XY, Kuo CCJ, Lew MS (eds) Advances in Multimedia Information Processing - PCM 2010. Lecture Notes in Computer Science, vol. 6298. Springer Berlin Heidelberg, pp 43–51
59. Xu M, Xu C, He X, Jin JS, Luo S, Rui Y (2013) Hierarchical affective content analysis in arousal and valence dimensions. Signal Process 93(8):2140–2150
60. Yang Y-H, Lin Y-C, Su Y-F, Chen HH (2007) Music emotion classification: a regression approach. In: IEEE international conference on multimedia and expo, 2007. IEEE, pp 208–211
61. Yang Y-H, Lin Y-C, Su Y-F, Chen HH (2008) A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process 16(2):448–457
62. Yang Yh, Chen HH (2012) Machine recognition of music emotion: a review. ACM Trans Intell Syst Technol 3(3):40
63. Yoo HW, Cho SB(2007) Video scene retrieval with interactive genetic algorithm. Multimed Tools Appl 34(3):317–336

64. Zhang M-L, Zhou Z-H (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
65. Zhang S, Tian Q, Jiang S, Huang Q, Gao W (2008) Affective mtv analysis based on arousal and valence features. In: IEEE international conference on multimedia and expo, 2008. IEEE, pp 1369–1372
66. Zhang S, Huang Q, Jiang S, Gao W, Tian Q (2010) Affective visualization and retrieval for music video. IEEE Trans Multimedia 12(6):510–522
67. Zhang M, Zhang K (2010) Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 999–1008

Shangfei Wang received the M.S. degree in circuits and systems, and the Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 1999 and 2002. From 2004 to 2005, she was a postdoctoral research fellow in Kyushu University, Japan. She is currently an Associate Professor of School of Computer Science and Technology, USTC. Dr. Wang is an IEEE member. Her research interests cover computation intelligence, affective computing, multimedia computing, information retrieval and artificial environment design. She has authored or coauthored over 50 publications.

Zhaoyu Wang received the Bachelor degree in School of Mathematics and Information Science from the Anhui University of Technology, Ma Anshan, Anhui Province, China, in 2010. He is studying for master degree in School of Computer Science and Technology from the University of Science and Technology of China, Hefei, Anhui Province, China. His research interest is Affective Computing.

Qiang Ji received his Ph.D degree in electrical engineering from the University of Washington. He is currently a Professor with the Department of Electrical, Computer, and Systems engineering at RPI. From January, 2009 to August, 2010, he served as a program director at the National Science Foundation, managing NSF's machine learning and computer vision programs. Prior to joining RPI in 2001, he was an assistant professor with Dept. of Computer Science, University of Nevada at Reno. He also held research and visiting positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, and the US Air Force Research Laboratory. Dr. Ji currently serves as the director of the Intelligent Systems Laboratory (ISL). Prof. Ji is a senior member of the IEEE. Prof. Ji's research interests includes computer vision, probabilistic graphical models, pattern recognition, information fusion for situation awareness and decision making under uncertainty, human computer interaction, and robotics.