Multimed Tools Appl (2015) 74:4679–4706
DOI 10.1007/s11042-013-1830-0

Implicit video emotion tagging from audiences' facial expression

**Shangfei Wang · Zhilei Liu · Yachen Zhu ·
Menghua He · Xiaoping Chen · Qiang Ji**

Published online: 12 January 2014

Abstract In this paper, we propose a novel implicit video emotion tagging approach by exploring the relationships between videos' common emotions, subjects' individualized emotions and subjects' outer facial expressions. First, head motion and face appearance features are extracted. Then, the spontaneous facial expressions of subjects are recognized by Bayesian networks. After that, the relationships between the outer facial expressions, the inner individualized emotions and the video's common emotions are captured by another Bayesian network, which can be used to infer the emotional tags of videos. To validate the effectiveness of our approach, an emotion tagging experiment is conducted on the NVIE database. The experimental results show that head motion features improve the performance of both facial expression recognition and emotion tagging, and that the captured relations between the outer facial expressions, the inner individualized emotions and the common emotions improve the performance of common and individualized emotion tagging.

S. Wang (✉) · Z. Liu · Y. Zhu · M. He · X. Chen
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, People's Republic of China, 230027
e-mail: sfwang@ustc.edu.cn

Z. Liu
e-mail: leivo@mail.ustc.edu.cn

Y. Zhu
e-mail: zhuyc@mail.ustc.edu.cn

M. He
e-mail: hemh@mail.ustc.edu.cn

X. Chen
e-mail: xpchen@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

Keywords Implicit emotion tagging · Expression recognition · Emotion inference

1 Introduction

Recent years have seen a rapid increase in the size of digital video collections. Because emotion is an important component in the human's classification and retrieval of digital videos, assigning emotional tags to videos has been an active research area in recent decades [35]. This tagging work is usually divided into two categories: explicit and implicit tagging [21]. Explicit tagging involves a user manually labeling a video's emotional content based on his/her visual examination of the video. Implicit tagging, on the other hand, refers to assigning tags to videos based on an automatic analysis of a user's spontaneous response while consuming the videos [21].

Although explicit tagging is a major method at present, it is time-consuming and brings users extra workload. However, implicit tagging labels videos based on the users' spontaneous nonverbal response while watching the videos. Therefore, implicit tagging can overcome the above limitations of the explicit tagging.

Since most of the current theories of emotion [13] agree that physiological activity is an important component of emotional experience, and several studies have demonstrated the existence of specific physiological patterns associated with basic emotions [25], recognizing subjects' emotion from physiological signals is one of the implicit video tagging methods [1, 29, 30]. There are many types of physiological signals, including Electroencephalography (EEG), Electrocardiography (ECG), Electromyography (EMG), and Galvanic skin resistance (GSR) etc. Present research has proved that physiological responses are potentially a valuable source of external user-based information for emotional video tagging. Physiological signals reflect unconscious changes in bodily functions, which are controlled by the Sympathetic Nervous System (SNS). These functions cannot be captured by other sensory channels or observer methods. However, physiological signals are susceptible to many artifacts, such as involuntary eye-movements, irregular muscle movements and environmental changes. These noises pose a significant challenge for signal processing and hinder the task of interpretation. In addition, subjects are required to wear complex apparatuses to obtain physiological signals, which may make some subjects feel uncomfortable.

Another implicit video tagging method is to recognize subjects' emotion from their spontaneous visual behavior [2, 3, 8, 22, 24], such as facial expressions, since recent findings indicate that emotions are primarily communicated through facial expressions and other facial cues (smiles, chuckles, frown, etc.). When obtaining an implicit tag by facial information, no complex apparatus other than one standard visible camera is needed. Thus this approach is more easily applied in real life. Furthermore, facial information is not significantly disturbed by body conditions, subjects can move their bodies as they wish. This freedom of motion makes them feel comfortable to express their emotions. Although spontaneous visual behavior is prone to environmental noise originating from lighting conditions, and occlusion, etc., it is more convenient and unobtrusive. Thus, implicit tagging using spontaneous behavior is a good and more practical alternative to neuro-physiological methods. Present research has already demonstrated that facial expressions can be a promising source to exploit for video emotion tagging. Most researchers have used the recognized expression directly as the emotional tag of the videos. However, although facial expressions are the major visual manifestation of inner emotions, they are not always consistent, which are two different concepts. In addition, facial expressions are more easier to be annotated than inner emotions. Extensive research in recent years on facial expression recognition has been conducted and much progress has been made in this area. Thus in this paper, we propose a new implicit tagging method by inferring subjects' inner emotions through a probabilistic model capturing the relations between outer facial expressions and the inner emotions, which is more feasible than the previous work that directly infers the inner emotion from the video/images or the methods that simply take the outer facial expressions as inner emotions. We assume the spontaneous facial expression reflects, to certain degree, the user's actual emotion as a result of watching a video. The expression hence positively correlates with user's emotion.

Furthermore, there are two kinds of emotional tags, the expected emotion and the actual emotion [7]. The expected emotion is contained in a video and intended to be communicated toward users from video program directors. It is likely to be elicited from majority of the users while watching that video. It can be considered as a common emotion. In contrast, the actual emotion is the affective response of a particular user to a video. It is context-dependent and subjective, and it may vary from one individual to another. It can be considered as an individualized emotion. Most present implicit tagging research has not considered both tags. In this research, we infer these two kinds of emotional tags and use both of them to tag videos.

Our tagging method consists of several steps. First, the eyes in the onset and apex expression images are located. Head motion features are computed from the coordinates of the eyes in the onset and apex frames, and face appearance features are extracted using the Active Appearance Model (AAM) [6]. Then, the subjects' spontaneous expressions are recognized using a set of binary Bayesian Network (BN) classifiers and Bayesian networks capturing the relations among appearance features (called structured BN) respectively. After that, the common emotional tags of videos are further inferred from the recognized expressions using a BN with three discrete nodes by considering the relations between the outer facial expressions, individualized emotions and the common emotions. The novelties of this work lie in the explicitly modeling the relationships between a video's emotional tag, the user's internal emotion, and the facial expression as well as in leveraging such relationships for more effective video emotion tagging. Through this model, we can indirectly infer a video's emotional content instead of directly treating subject's expression as the video emotional tag as being done by the existing implicit video emotion tagging methods.

The outline of this paper is as follows. First, in Section 2 we introduce previous work related to implicit emotion tagging by using physiological signals and spontaneous behaviors. Then, our proposed implicit emotion tagging approach is explained in detail in Section 3. The experiments and analyses of facial expression recognition and emotion tagging are described in Section 4. Finally, some discussions and conclusions are presented in Section 5.

2 Related work

An increasing number of researchers have studied emotional video tagging from subjects' spontaneous responses. Vinciarelli et al. [21] is the first to present the disadvantages of explicit tagging and introduce the concept, implementation and main problems of implicit Human-Centered tagging. Currently, implicit emotion tagging of videos mainly uses physiological signals or subjects' spontaneous visual behavior. In this section, the related studies are briefly reviewed.

2.1 Affective video content analyses using physiological signals

Several researchers have focused on implicit tagging using physiological signals, which could reflect subtle variations in the human body. Money et al. [17, 18] investigated whether users' physiological responses, such as galvanic skin response (GSR), respiration, Blood Volume Pulse (BVP), Heart Rate (HR) and Skin Temperature (ST), can serve as summaries of affective video content. They collected 10 subjects' physiological responses during watching three films and two award-winning TV shows. Experimental results showed the potential of the physiological signals as external user-based information for affective video content summaries. They [19] further proposed Entertainment-Led Video Summaries (ELVIS) to identify the most entertaining sub-segments of videos based on their previous study.

Soleymani et al. [29, 30] analyzed the relationships between subjects' physiological responses, the subject's emotional valence as well as arousal, and the emotional content of the videos. A dataset of 64 different scenes from eight movies were shown to eight participants. The experimental results demonstrated that besides multimedia features, subjects' physiological responses (such as GSR, EMG, blood pressure, respiration and ST) could be used to rank video scenes according to their emotional content. Moreover, they [9] implemented an affect-based multimedia retrieval system by using both implicit and explicit tagging methods. Soleymani et al. further [11] constructed two multimodal datasets for implicit tagging. One is DEAP (Database for Emotion Analysis using Physiological Signals) [11], in which EEG and peripheral physiological signals, including GSR, respiration, ST, ECG, BVP, EMG and electrooculogram (EOG), were collected from 32 participants during their watching of 40 one-minute long excerpts of music videos. Frontal face videos were also recorded from 22 among 32 participants. The other database is MAHNOB-HCI [32], in which face videos, audio signals, eye gaze, and peripheral/central nervous system physiological signals of 27 subjects were recorded during two experiments. In the first experiment, subjects selfreported their felt emotions to 20 emotion-induced videos using arousal, valence, dominance and predictability as well as emotional keywords. In the second experiment, subjects assessed agreement or disagreement of the displayed tags with the short videos or images.

While these two pioneer groups investigated many kinds of physiological signals as the implicit feedback, other researchers focused only on one or two kinds of physiological signals. For example, Canini et al. [4] investigated the relationship between GSR and affective video features for the arousal dimension. Using correlation analysis on a dataset of 8 subjects watching 4 video clips, they found a certain dynamic correlation between arousal, derived from measures of GSR during film viewing, and specific multimedia features in both audio and video domains. Smeaton et al. [27] proposed to detect film highlights from viewers' HR and GSR. By comparing the physiological peaks and the emotional tags of films on a database of 6 films viewed by 16 participants, they concluded that subjects' physiological peaks and emotional tags are highly correlated and that music-rich segments of a film do act as a catalyst in stimulating viewer response. Toyosawa et al. [33] proposed to extract attentive shots with the help of subjects' heart rate and heart rate variability.

Two researcher groups considered event-related potential (ERP) as subjects' implicit feedback. One [10] attempted to validate video tags using an N400 ERP on a dataset with 17 subjects, each recording for 98 trials. The experimental results showed a significant difference in N400 activation between matching and non-matching tag. Koelstra et al. [12]

also found robust correlations between arousal and valence and the frequency powers of EEG activity. The other [36] attempted to perform implicit emotion multi-media tagging through a brain-computer interface system based on a P300 ERP. 24 video clips (four clips were chosen for each of the six basic emotional categories (i.e. joy, sadness, surprise, disgust, fear, and anger)) and 6 basic facial expression images were displayed to eight subjects. The experimental results showed that their system can successfully perform implicit emotion tagging and naive subjects who have not participated in training phase can also use it efficiently.

Instead of using contact and intrusive physiological signals, Krzywicki et al. [14] adopted facial thermal signatures, a nonconstant and non-intrusive physiological signal to analyze affective content of films. They examined the relationship between facial thermal signatures and emotion-eliciting video clips on a dataset of 10 subjects viewing three film clips that were selected to elicit sadness and anger. By comparing the distribution of temperatures with the summarized video clip events, they concluded that changes in the global temperature are consistent with changes in stimuli and that different regions exhibit different thermal pattern in response to stimuli.

Other than analyzing affective content of videos, Arapakis et al. [1] predicted the topic relevance between query and retrieved results by analyzing implicit feedback, which includes facial expression and peripheral physiological signals such as GSR and ST. Their results showed that the prediction of topic relevance is feasible, and the implicit feedback can benefit from the incorporation of affective features.

These studies described above have indicated the potential of using physiological signals for the implicit emotion tagging of videos. However, to acquire physiological signals, subjects are required to wear several contact apparatuses, which may make them feel uncomfortable and hinder the real application of this method. Furthermore, some research also indicates that the accuracy of current emotion detection method from physiological signals is not superior to multimedia content analysis or high enough to replace the self-reports [31]. The improvement or new methods are needed to meet the requirement in a real application.

2.2 Affective video content analyses using spontaneous visual behavior

Several researchers have turned to affective video content analyses according to human spontaneous visual behavior, since it can be measured using non-contact and non-intrusive techniques, and easily applied in real life. Hideo Joho et al. [3, 8] proposed to detect personal highlights in videos by analyzing viewers' facial activities. The experimental results on a dataset of 10 participants watching eight video clips suggested that compared with the activity in the lower part, the activity in the upper part of face tended to be more indicative of personal highlights.

Peng et al. [22, 24] proposed to fuse users' eye movements (like blink or saccade) and facial expressions (positive or negative) for home video summarization. Their experimental results on 8 subjects watching 5 video clips, demonstrated the feasibility of both eye movements and facial expressions for video summarization application. They [23] also proposed and integrated an interest meter module into a video summarization system, and achieved good performance.

Liu et al. [15] proposed an implicit video multiple emotion tagging method by exploiting the relations among multiple expressions, and the relations between outer expressions and inner emotions. The experimental results on the NVIE database demonstrated that multi-expression recognition considering the relations among expressions improved the recognition performance. The tagging performance considering the relations between expression and emotion outperformed the traditional expression-based implicit video emotion tagging methods.

Other than focusing on subjects' whole facial activity, Kok-Meng Ong [20] analyzed affective video content from viewers' pupil sizes and gazing points. Experimental results on 6 subjects watching 3 videos showed the effectiveness of their approach.

Instead of affective content analysis of videos, Ioannis Arapakis et al. [2, 3] proposed a video search interface that predicts the topical relevance by incorporating facial expressions and click-through action into user profiling and facilitating the generation of meaningful recommendations of unseen videos. The experiment on 24 subjects demonstrated the potential of multi-modal interaction for improving the performance of recommendation.

Although all the studies described above explored visual behavior to analyze the affective content of a video, their purposes (i.e. summarization [8, 22, 24], recommendation [2, 3], tagging [15]) and the used modalities (i.e. facial expression [2, 3, 8, 22, 24], click-through action [2, 3], eye movements [3, 22, 24]) are not the same. The facial expression classifiers used in the related work are eMotion (a facial expression recognition software) [2, 3], Support Vector Machine (SVM) [22, 24] and Bayesian networks [8, 15].

These studies illustrate the development of methods for using spontaneous visual behavior in the implicit tagging of videos. However, the assumptions made by the above studies is that the expressions displayed by the subjects were the same as their internal feelings when they watched the videos. For this reason, most researchers have used the recognized expression directly as the emotional tag of the videos. However, research has indicated that internal feelings and displayed facial behaviors are related, but not always the same [5] because some emotions are not always expressed in our daily life. Furthermore, few research has paid attention to common emotion and individualized emotion. Therefore, in this paper we propose emotion tagging of videos by inferring videos' common emotions and users' individualized emotions from users' expressions. Furthermore, the data sets used in previous studies were small, with the number of the subjects ranging from 6 to 32. Thus, a much larger emotion database named NVIE database [34] is constructed, in which the facial videos of 128 subjects were recorded when they watched emotional videos in three types of illumination conditions (i.e., front, left and right).

Compared with the most related work [15], we find that paper [15] focuses on modeling the co-occurrence and mutually exclusive relationships among different facial expressions for improved video emotion tagging. It does not differentiate the subject's internal emotion from the video common emotion, treating them as the same. This paper, on the other hand, explicitly models the differences and the relationships between a video's common emotion and the user's internal emotion. These two works, hence, addressed different problems.

3 Implicit emotion tagging approach

Figure 1 gives the overview of our implicit emotion tagging approach. It consists of two components: expression recognition model and video emotion tagging model

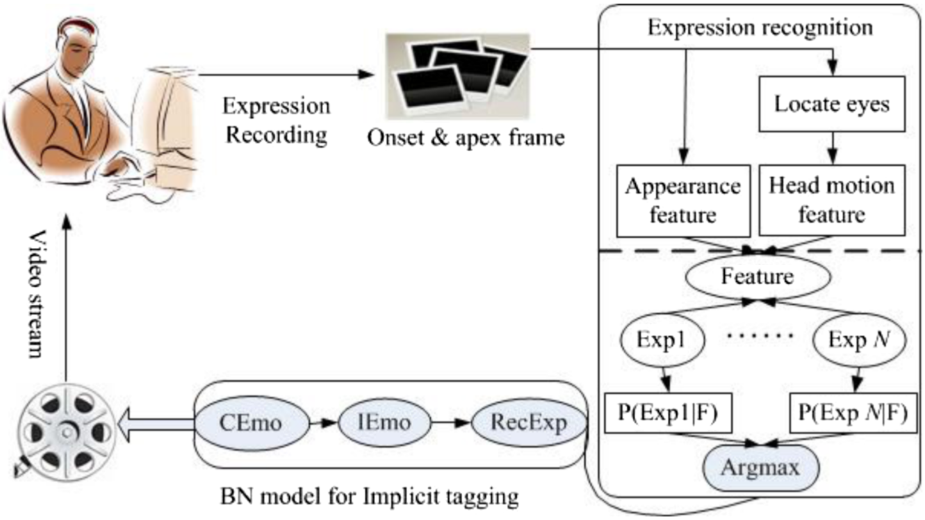


Fig. 1 Framework of our method

based on recognized facial expressions. Details for each component are discussed below.

3.1 Facial expression recognition

Facial expression recognition includes facial feature extraction, feature selection, and expression recognition.

3.1.1 Facial feature extraction

Two kinds of features are extracted: head motion features and facial appearance features. Two head motion features, including the translational speed of head motion and the head's rotational speed, are calculated. First, the subject's eyes are located automatically by using eye location method based on AdaBoost and Haar features [16]. Then, head motion features are calculated from the coordinates of the eyes in the onset and apex frames as follows:

$$Speed_m = \frac{\sqrt{\left(C_x^{apex} - C_x^{onset}\right)^2 + \left(C_y^{apex} - C_y^{onset}\right)^2}}{Time} \tag{1}$$

$$Speed_r = \frac{\left|\arctan\left(\frac{R_y^{apex} - L_y^{apex}}{R_x^{apex} - L_x^{apex}}\right) - \arctan\left(\frac{R_y^{onset} - L_y^{onset}}{R_x^{onset} - L_x^{onset}}\right)\right|}{Time} \tag{2}$$

where (C_x, C_y) represents the coordinate of the center of the two eyes, $Time = frame_{apex} - frame_{onset}$, and (L_x, L_y) and (R_x, R_y) represent respectively the coordinates of the left and right eye locations. In (1), $Speed_m$ is the speed of the head motion. In (2), $Speed_r$ represents the rotational speed of the head. $Speed_m$ and $Speed_r$ are both scalars.

Besides motion features, facial appearance features are also extracted. Since the AAM captures information about both appearance and shape [6], we use AAM to extract visible features from the apex expressional images. The AAM tools from [26] are used to extract the AAM features here.

All apex images were rotated to arrange the two eyes in a horizontal line and then normalized to 400 × 400 grayscale images with the center of the two eyes at (200, 160). The face was labeled with 61 points as shown in Fig. 2. One third of the apex images were selected to build the appearance model. This model was then applied to the remaining images to obtain their appearance parameters as the appearance feature. Here, AAMs are trained in a person-independent manner. Finally, a 30-dimension feature vector was extracted from each of the apex images using the AAM algorithm.

3.1.2 Feature selection

In order to select distinctive features for each Naive BN classifier of each expression category, the F-test statistic [37] is used for feature selection. Like the Fisher criterion, F-statistic is the ratio of between group variance to within-group-variance. The significance of all

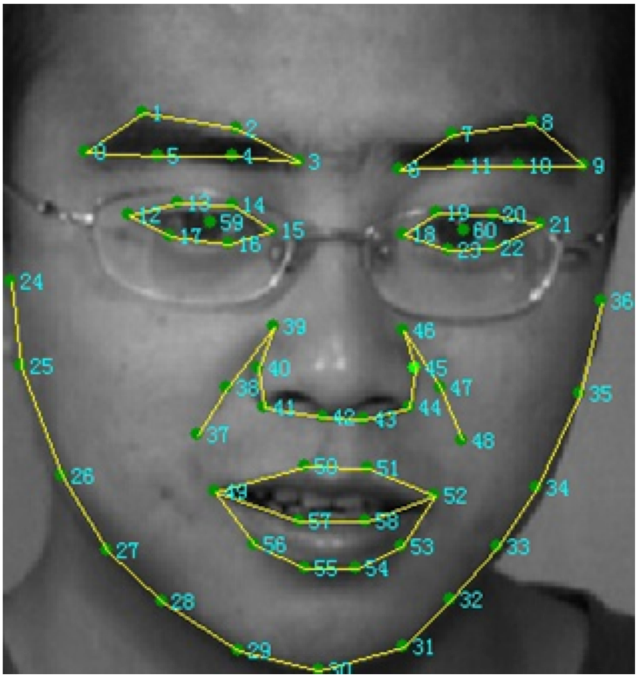


Fig. 2 Distribution of AAM points

features can be ranked by sorting their corresponding F-test statistics in descending order. The F-test statistic of feature X is calculated by using the following equation:

$$F(X) = \left(\frac{\sum_{c=1}^{N} n_c (\overline{x_c} - \overline{x})^2}{\sum_{c=1}^{N} (n_c - 1)\sigma_c^2} \right) \left(\frac{n - N}{N - 1} \right) \tag{3}$$

where N is the number of classes, n_c is the number of samples of class c, n is total number of samples, $\overline{x_c}$ is the average of feature X within class c, $\overline{x}$ is the global average of feature X, and σ_c^2 is the variance within class c. According to the calculated F-statistics of all features, features are selected from high to low.

3.1.3 Expression recognition

Expression recognition through Naive BNs Given the selected facial features, we propose to use naive BNs to recognize facial expression due to its simplicity. BN is a probabilistic graphical model (PGM) that encodes the causal probabilistic relationships of a set of random variables via a directed acyclic graph (DAG), where the nodes represent the random variables and the edges represent the conditional dependencies between variables. Compared with other commonly used deterministic classifiers such as the SVM, BN is simple and can effectively model the vagueness and uncertainties with the affective states and facial features. In addition, BN offers principled inference method to perform classification.

In order to select discriminative features for each kind of expression, we construct N binary BNs instead of one multi-class BN. It means that, the N-class expression recognition problem is solved by N BN classifiers for each kind of facial expressions as shown in Fig. 3a.

Each BN consists two nodes, feature node F and category node C as shown in Fig. 3b. The former is a continuous node, and the latter is a discrete node with two states (1 and 0) representing the recognition result being expression C_i and not C_i respectively. Given the BN's structure, the BN parameters, i.e., the prior probability of C, P(C), and the conditional probability, $P(F|C)$, are learnt from the training data though maximum likelihood (*ML*) estimation. After training, the posterior probability $P(C_i = 1|F)$ of a testing sample is calculated according to (4):

$$P(C_i = 1|F) = \frac{P(C_i = 1, F)}{P(F)} = \frac{P(F|C_i = 1)P(C_i = 1)}{P(F)} \tag{4}$$

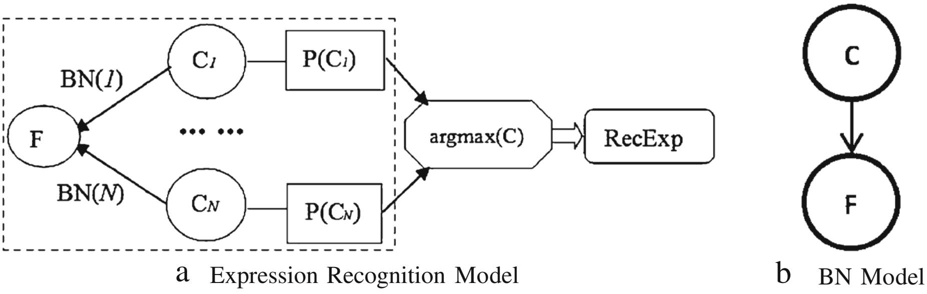


Fig. 3 Expression recognition model and a simple BN model

After all the posterior probabilities for each expression have been calculated, the final recognized expression can be obtained as follows:

$$RecExp^* = \arg\max_{C} P(C_i = 1|F) \tag{5}$$

Expression recognition through modeling the structure of the feature points using BN Instead of using Naive BN, we propose another sets of BN to capture the structure of the feature points embedded in N-class expressions (called structured BN), as shown in in Fig. 4. Each node of the BN is a geometric feature (i.e. the coordinates of feature points and the head motions), and the links and their conditional probabilities capture the probabilistic dependencies among all the geometric features.

The BN learning consists of structure learning and parameter 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. 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:

$$Score(G) = \max_{\theta} log(p(DL|G, \theta)) - \frac{Dim_G}{2} logm \tag{6}$$

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

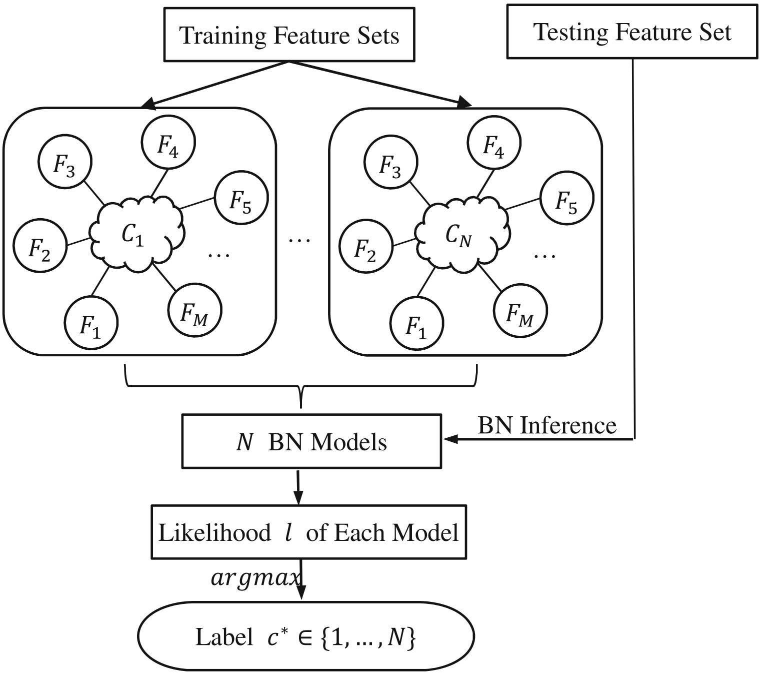


Fig. 4 Expression recognition through modeling their geometric features

parameters. After the BN structure is constructed, parameters can be learned from the training data. Because a complete training data is provided in this work, Maximum Likelihood Estimation (MLE) method is used to estimate the parameters.

In this work, N models $\Theta_c, c = 1, \cdots, N$ are established during training, where N is the number of expression categories. After training, the learned BNs capture the muscle movement pattern for N-class expressions respectively.

During testing, the samples are classified into the cth expression according to

$$\begin{aligned} c^{\star} &= \arg\max_{c\in[1,n]} \frac{P(E_T|\Theta_c)}{Complexity(M_c)} \\ &= \arg\max_{c\in[1,n]} \frac{\prod_{i=1}^{M} P_c(F_i|pa(F_i))}{Complexity(M_c)} \\ &\propto \arg\max_{c\in[1,n]} \sum_{i=1}^{M} log(P_c(F_i|pa(F_i))) - log(Complexity(M_c)) \end{aligned} \tag{7}$$

where E_T represents the features of a sample, $P(E_T|\Theta_c)$ denotes the likelihood of the sample given the cth model, M represents the dimensions of the features that is the number of nodes, F_i is the ith node in the BN, and $pa(F_i)$ denotes the parent nodes of F_i, and M_c stands for cth model and $Complexity(M_c)$ represents the complexity of M_c. Since different models may have different spatial structures, the model likelihood $P(E_T|\Theta_c)$ will be divided by the model complexity for balance. We use the total number of the links as the model complexity.

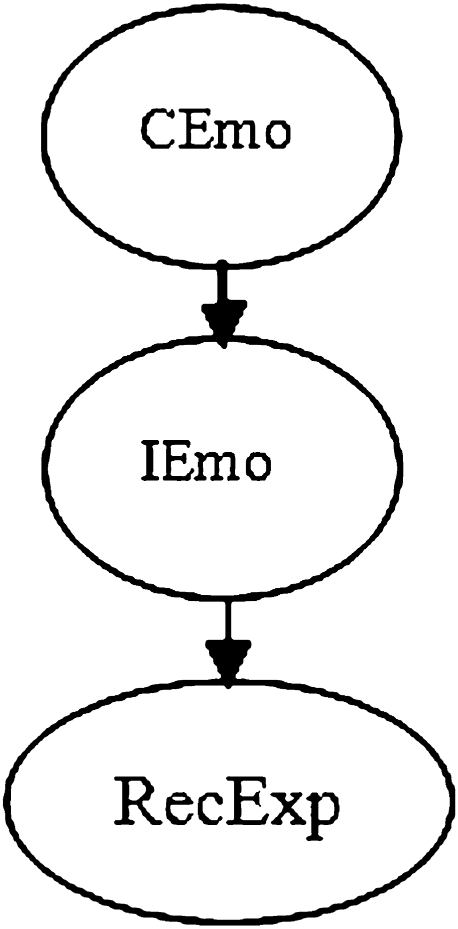
3.2 Emotion tagging of videos

The emotion elicitation process can be captured by a generative process using a video to induce user's emotion, which, in turn, causes certain facial expression on the user's face as an external manifestation of the user's internal emotion. To model this generative process and to capture the inherent causal relationships between a video's emotion content, user's internal emotion, and user facial expression, we propose to use another BN, as shown in Fig. 5, for video emotion tagging.

As a graphical model, BN can effectively captures the causal relationships among the random variables. It is therefore a natural choice to capture and model the natural and inherent relationships between the common emotion of the video, the specific emotion of the user, and the user's facial expression. Moreover, BN also allows rigorous inference of video emotion tag from the recognized facial expressions. This BN includes three discrete nodes and links. The nodes respectively represent the common emotion tag (CEmo), the individualized emotion tag (IEmo) and the recognized expression (RecExp), while the links capture the causal relationships among the nodes. Each node has N states, representing N classes.

Given the BN in Fig. 5, a similar maximum likelihood estimation process is used to estimate its parameters, i.e., the conditional probabilities of each node including $P(CEmo)$ (the prior probability of the common emotion), $P(IEmo|CEmo)$ (the conditional probabilities of the subjects' individualized emotion state given the video's common emotion tag) and $P(RecExp|IEmo)$ (the conditional probability of the training sample's recognized expression state given its individualized emotion state). During testing, the posterior probabilities

Fig. 5 Common **emotion** (CEmo) recognition based on recognized expressions (RecExp) and subjects' individualized emotions (IEmo)



of the video's individualized emotion tag $IEmo^*$ and common emotion tag $CEmo^*$ are inferred using the following equations:

$$\begin{aligned} IEmo^* &= \arg\max_{IEmo} P(IEmo|RecExp) \\ &= \arg\max_{IEmo} \sum_{CEmo} P(CEmo)P(IEmo|CEmo)P(RecExp|IEmo) \end{aligned} \tag{8}$$

$$\begin{aligned} CEmo^* &= \arg\max_{CEmo} P(CEmo|RecExp) \\ &= \arg\max_{CEmo} \sum_{IEmo} P(CEmo)P(IEmo|CEmo)P(RecExp|IEmo) \end{aligned} \tag{9}$$

4 Implicit emotion tagging experiments

4.1 Databases for implicit video tagging

As mentioned in Section 2.1, Soleymani et al. constructed two multimodal datasets for implicit tagging, DEAP and MAHNOB-HCI. Both consist of the facial images of subjects when they watch videos. However, neither of them has facial expression annotation. Thus, due to the fact that facial expression annotation is both onerous and subjective, and that we do not have the necessary expertise to do an objective expression annotation at present, we don't use these two databases in this work. The NVIE (Natural Visible and Infrared facial Expression) database [34] is another multimodal database for facial expression recognition and emotion inference. NVIE contains both posed expressions and video-elicited spontaneous expressions of more than 100 subjects under three different illumination directions. During the spontaneous expression collection experiments, the participants offered the self-report to the stimuli video according to their emotion experiences in six basic emotion categories, named happiness, disgust, fear, sadness, surprise and anger, which can be regarded as the individualized emotional tags. The common emotional tags are determined by a majority rule. In addition, the NVIE database provides the facial expression annotations of both apex facial images and image sequences in six categories. This database is therefore suitable for our implicit video emotion tagging experiments. The database consists the samples where the expression positively correlates with user's emotion, does not contain cases where the user's expression negatively correlates with their actual emotion. The construction details of the NVIE database can be found in [34]. Appendix presents the information of stimulus videos and subjects.

For the purpose of this study, the facial image sequences whose emotion categories and expression categories are happiness, disgust, fear, surprise, sadness and anger, and whose average evaluation values of the self-report data are larger than 1 are selected from the NVIE database. Thus, six expression and emotion categories are considered in this paper. Ultimately, we selected 1154 samples and the annotations of their expressions, individualized emotions and videos' common emotional tags as shown in Table 1, in which a total of 32 videos (including 6 happiness videos, 6 disgust videos, 5 fear videos, 7 surprise videos, 4 anger videos, 4 sadness videos) are watched by these subjects. Besides, the confusion relations between the subjects' expressions, individualized emotional tags, and common emotional tags are summarized in Table 2.

From Table 2, it is clear that, although there are high consistencies between facial **expressions**, individualized **emotions**, and common emotions, there still exist some discrepancies, especially for some negative emotion or expression states such as anger. This suggests that while outer facial expressions can well reflect our inner individualized or common emotions, they are not completely the same. Furthermore, the table also shows the differences between a subject's individual emotion and the video's common emotion. This means the

Table 1 The information of the selected samples

Num.	Hap.	Dis.	Fear	Sur.	Ang.	Sad.
Exp	326	222	163	162	156	125
IEmo	300	235	167	197	133	122
CEmo	287	212	168	193	152	142

Table 2 Relations between the samples' expressions, individualized emotions and common emotions

CEmo	Average: 0.8709					
Exp	Hap.	Disg.	Fear	Surp.	Ang.	Sad.
Hap.	282	4	1	33	6	0
Disg.	0	188	10	8	13	3
Fear	0	9	150	1	1	2
Surp.	3	3	5	150	0	1
Ang.	1	5	1	0	124	25
Sad.	1	3	1	1	8	111
CEmo	Average: 0.8666					
IEmo	Hap.	Disg.	Fear	Surp.	Ang.	Sad.
Hap.	280	3	0	15	2	0
Disg.	3	174	20	2	35	1
Fear	0	23	142	0	2	0
Surp.	4	9	6	176	0	2
Ang.	0	0	0	0	111	22
Sad.	0	3	0	0	2	117
IEmo	Average: 0.7842					
Exp	Hap.	Disg.	Fear	Surp.	Ang.	Sad.
Hap.	282	6	1	33	4	0
Disg.	3	160	30	12	11	6
Fear	0	26	127	7	2	1
Surp.	12	3	5	141	0	1
Ang.	2	35	3	0	99	17
Sad.	1	5	1	4	18	96

same video with the same common tag may invoke different individual emotions for different people. If we can exploit these relationships effectively, it may be helpful in emotion reasoning or video tagging from facial expressions.

4.2 Experimental conditions

To select the best dimension of features for the naive BN classifiers, we employ a model selection strategy and 10-fold cross validation. First, all the samples are divided into ten parts. One of them is used as the test set, and the remaining are used as the training set. We apply a 10-fold cross validation on the training set to choose the features that achieve the highest accuracy rate on the validation set. After that, the selected features and the constructed BNs are used on the test set to classify facial expressions.

In order to evaluate the effectiveness of our approach from different aspects, two commonly used parameters, the precision and the F_1 score [28], are adopted, which are defined as follows:

$$Precision(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)} \tag{10}$$

$$F_1(C_i) = \frac{2 \times TP(C_i)}{2 \times TP(C_i) + FN(C_i) + FP(C_i)} \tag{11}$$

where, TP (true positive) represents the number of samples correctly labeled as belonging to the positive class C_i, FP (false positive) is the number of samples incorrectly labeled as belonging to the positive class C_i, and FN (false negative) is the number of the samples which are not labeled as belonging to the positive class C_i but should have been.

Our work focuses on emotional tagging using facial expressions, while the current related works mentioned in Section 2.2, explored facial expression, click-through action, or eye movements for video summarization, recommendation and tagging. The purposes and modalities of related works are not exactly the same as our work. Therefore, we cannot directly compare our work with these works. Through analyses, we find that the facial expression classifiers used in the related work are eMotion (a facial expression recognition software) [2, 3], SVM [22, 24] and Bayesian networks [8, 15]. The Bayesian networks used in [8] are similar to our structured BN. Therefore, the experimental results using structured BN can be regarded as the comparison with [8].

4.3 Experimental results and analyses

4.3.1 Experimental results and analyses of expression recognition

According to the expression classification model described in Section 3.1.3, two comparative experiments using only the AAM features and the combination of AAM features and head motion features are performed to recognize the outer expressions. The classification precisions and the F_1 scores of the two experiments are shown in Table 3.

From Table 3, we can find the following two phenomenons: (1) For both expression recognition methods, the overall precisions by using the AAM features and AAM+DHM features are 0.602 vs. 0.640 and 0.530 vs. 0.563 respectively. It means that head motion features improve the overall classification results more than 3 %. The average F_1 scores of the classifiers with head motion are also higher than those without head motion, which indicates that the head motion is useful for spontaneous facial expression recognition. (2) The recognition rate of happiness is high, and the recognition rates of the negative expressions are relatively low. The reason may be that, it is easier to elicit the positive expressions than negative expressions by using the video-based emotion elicitation method [5].

For the Naive BN classifiers, the selection probabilities of all the 32 features including the head motion features (feature ID: 31–32) over the ten folds are shown in Fig. 6. From Fig. 6, we can conclude that: (1) For happiness, disgust, and fear, the head motion features are selected. It means that the head motion features are helpful for distinguishing these

Table 3 Expression recognition results with (AAM+HM) and without (AAM) head motion features

Method	Feature	Measure	Hap.	Dis.	Fear	Sur.	Ang.	Sad.	Average
Naive BN	AAM	Precision	0.896	0.527	0.092	0.747	0.526	0.544	0.602
		F_1 score	0.845	0.557	0.154	0.564	0.492	0.567	0.530
	AAM+HM	Precision	0.917	0.572	0.384	0.648	0.455	0.584	0.640
		F_1 score	0.841	0.579	0.502	0.577	0.481	0.589	0.595
Structured BN	AAM	Precision	0.801	0.383	0.178	0.630	0.513	0.672	0.530
		F_1 score	0.821	0.466	0.220	0.534	0.457	0.540	0.506
	AAM+HM	Precision	0.785	0.356	0.380	0.679	0.500	0.680	0.563
		F_1 score	0.804	0.438	0.486	0.563	0.458	0.526	0.546

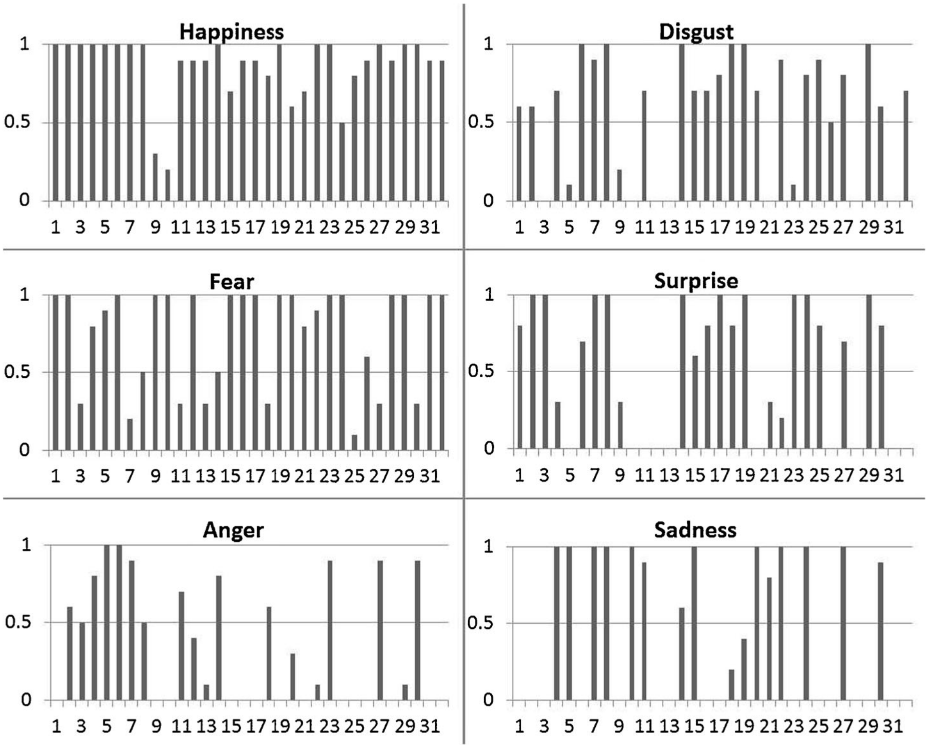


Fig. 6 Feature selection results of expression recognition using Naive BN

expressions when naive BNs are used. This conclusion can also be reflected in the recognition results given in Table 3. (2) The selected feature numbers for different expressions are different. For happiness, disgust, and fear, more than half of the features are selected, while for the other three expressions, only a few features are selected, especially for anger and sadness. This proves that the discriminative features for different expressions are different.

For the structured BN, the learned BNs are showed in Fig. 7. From the figure, we can find that: (1) The learned structure for the six expressions are different. It may indicate that the appearance features' relations embedded in different expressions are different. (2) For most expressions, one or two head motion features are dependent with AAM features. It may confirm that, head motions are related to facial appearance for expression manifestations.

4.3.2 Experimental results and analyses of emotion tagging

Based on the recognized facial expressions, the subjects' individualized emotion states as well as the video's common emotional tags are inferred by a 3-node BN model. Tables 4 and 5 present the precisions and F_1 scores of tagging results. From the tables, we can find that both the precisions and F_1 score of the individualized tagging are lower than those of the common tagging. It illustrates the difficulty with personalized video emotion tagging, since individualized emotions are context dependent, subjective and complex.

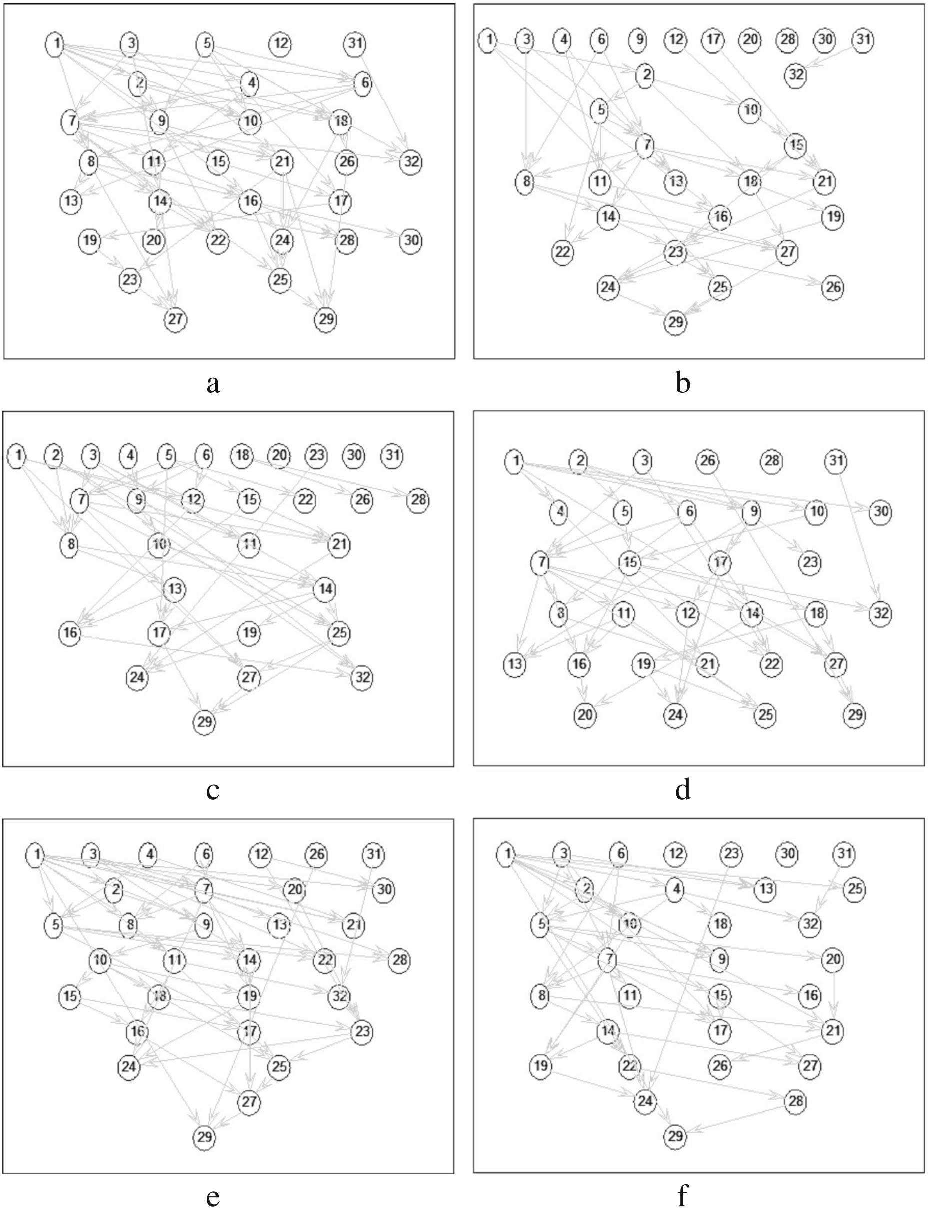


Fig. 7 The learned BN structures for six expressions using AAM and head motion features; **a** Happiness; **b** Disgust; **c** Fear; **d** Surprise; **e** Anger; **f** Sadness

To further validate the effectiveness of our proposed tagging method, comparative tagging experiments, which recognize the common and individualized emotional tags directly from the facial features, are conducted. The classification models are similar to the expression classification models described in Section 3.1.3, where the original expression labels of the samples are replaced by the common and individualized emotional tags.

Table 4 Individualized emotion tagging results based on AAM and AAM+HM features

Tagging method	Feature	Tagging phase	Measure	Hap.	Dis.	Fear	Sur.	Ang.	Sad.	Average
Naive BN	AAM	Fea. –>IEmo	Precision	0.850	0.434	0.126	0.548	0.459	0.590	0.536
			F_1 score	0.801	0.468	0.179	0.481	0.427	0.543	0.483
		Exp. –>IEmo	Precision	0.877	0.451	0.084	0.594	0.504	0.534	0.548
			F_1 score	0.791	0.490	0.141	0.504	0.432	0.549	0.484
	AAM+HM	Fea. –>IEmo	Precision	0.860	0.413	0.228	0.599	0.429	0.566	0.552
			F_1 score	0.778	0.462	0.321	0.505	0.433	0.535	0.506
		Exp. –>IEmo	Precision	0.893	0.477	0.287	0.513	0.444	0.557	0.569
			F_1 score	0.782	0.496	0.376	0.506	0.434	0.555	0.525
Structured BN	AAM	Fea. –>IEmo	Precision	0.740	0.328	0.252	0.492	0.248	0.664	0.454
			F_1 score	0.774	0.404	0.290	0.473	0.222	0.455	0.436
		Exp. –>IEmo	Precision	0.800	0.336	0.198	0.513	0.489	0.640	0.496
			F_1 score	0.787	0.418	0.246	0.484	0.398	0.507	0.473
	AAM+HM	Fea. –>IEmo	Precision	0.757	0.315	0.281	0.564	0.165	0.689	0.462
			F_1 score	0.775	0.395	0.375	0.514	0.154	0.444	0.443
		Exp. –>IEmo	Precision	0.790	0.298	0.299	0.569	0.459	0.631	0.508
			F_1 score	0.776	0.374	0.386	0.526	0.384	0.481	0.488

Table 5 Common emotion tagging results based on AAM and AAM+HM features

Tagging method	Feature	Tagging phase	Measure	Hap.	Dis.	Fear	Sur.	Ang.	Sad.	Average
Naive BN	AAM	Fea. –>CEmo	Precision	0.868	0.509	0.161	0.627	0.270	0.669	0.556
			F_1 score	0.819	0.529	0.220	0.513	0.328	0.586	0.499
		IEmo –>CEmo	Precision	0.909	0.542	0.089	0.627	0.487	0.500	0.569
			F_1 score	0.801	0.561	0.150	0.526	0.450	0.553	0.507
	AAM+HM	Fea. –>CEmo	Precision	0.861	0.575	0.345	0.622	0.237	0.655	0.586
			F_1 score	0.795	0.560	0.458	0.543	0.308	0.578	0.540
		IEmo –>CEmo	Precision	0.923	0.580	0.345	0.544	0.421	0.542	0.600
			F_1 score	0.789	0.573	0.453	0.532	0.440	0.581	0.561
Structured BN	AAM	Fea. –>CEmo	Precision	0.770	0.382	0.268	0.523	0.296	0.683	0.487
			F_1 score	0.788	0.453	0.309	0.498	0.285	0.516	0.475
		IEmo –>CEmo	Precision	0.829	0.401	0.208	0.534	0.467	0.627	0.511
			F_1 score	0.797	0.479	0.260	0.499	0.410	0.543	0.498
	AAM+HM	Fea. –>CEmo	Precision	0.788	0.359	0.351	0.596	0.237	0.711	0.507
			F_1 score	0.788	0.434	0.468	0.535	0.234	0.513	0.495
		IEmo –>CEmo	Precision	0.819	0.354	0.357	0.596	0.461	0.634	0.537
			F_1 score	0.786	0.427	0.462	0.545	0.415	0.529	0.528

The precisions and F_1 scores of the experiment results are shown in Tables 4 and 5. Comparisons and corresponding conclusions are listed as follows:

- By comparing the results of directly inferring tag from image features to our method, we can find that the video's common emotion tagging and subjects' individualized emotion recognition results when considering relations among the outer expression, individualized emotional tags and the videos' common emotional tags are superior to those without considering these relationships in terms of both precision and F_1 score. It proves the effectiveness of our proposed implicit emotion tagging method.
- Comparing the results of using and without using head motion features, it is clear that head motion features improve the overall tagging results in terms of both the precision and the F_1 score. As for the specific categories, head motion features can improve the precision of happiness, disgust, sadness and especially fear.

5 Conclusions

Emotion tagging of videos has been an active research area in recent decades. Implicit tagging using audiences' spontaneous response has become increasingly attractive, and preliminary research has been performed because of its potential applications. In this paper, we propose an implicit video tagging method from the subjects' spontaneous facial expression. To recognize facial expressions, a set of binary Naive BNs and structured BNs are employed. The common and individual emotional tags of a video are then inferred from the recognized facial expressions through a 3-node BN by explicitly modeling the relations among the outer facial expressions, the individualized emotional tags and the common emotional tags. The results show that head motion features improve the overall performance of the spontaneous expression recognition, the subjects' individualized emotion tagging and the videos' common emotion tagging. The captured relations among the outer facial expressions, individualized emotional tags and the common emotional tags are helpful for implicit video tagging. However, the performance improvement is minor and incremental. This may be due to the fact that the relationships vary with subject and with emotion. We shall further investigate this issue in the future. Our method requires the dataset must have simultaneous annotations for facial expression, audiences' inner emotion and videos emotion tags. This, unfortunately, is not the case for the existing databases except NVIE database. For example, both DEAP and MAHNOB-HCI databases do not have facial expression annotations. To use these databases or any existing databases requires us to provide the missing annotations. Annotation of any database is an onerous and time-consuming task. Furthermore, annotation requires the necessary expertise in order to provide accurate and objective labels. We currently do not have such an expertise. Thus, in this paper, we only evaluate our method on NVIE database. We will perform further evaluation on another database in the future.

The existing implicit tagging work regards the subject's facial expressions as the video's emotional tags directly, and has rarely considered both individualized emotional tag and common emotional tag. Compared with these work, we are the first to propose a BN classifier to systematically capture the differences and relations among the outer facial expressions, subjects' inner individualized emotion states, and the videos' common emotion categories. We find that the emotion tagging results based on facial expression recognition and BN inference considering these three items' relations are better than the results of direct individualized emotion tagging or videos' common emotion tagging from facial images.

Acknowledgments 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.

Appendix

A total of 32 playlists including six basic emotional video clips are prepared for the subjects. The numbers of subjects of different playlists under three illuminations are shown in Table 6.

Playlists 31 and 32 were used for supplementary experiments (when a subject's emotions were not elicited successfully, another experiment was conducted using these two playlists). Playlist 13 was used only as a complement to induce three emotions: disgust, anger, and happiness. Table A-1 shows that 56 subjects used playlist 11; however, their last three emotions (disgust, anger, and happiness) were induced by playlist 13 again. This was performed because for some of the early experiments the playlists were too long to record with the camera; therefore, a supplementary experiment was carried out for these subjects using a playlist including the last three emotion-eliciting videos. The contents of these playlists are shown in Table 7. All of these video clips are segmented from some movies or TV shows obtained from the internet. A brief description of each video's content is provided below.

Happy-1.flv: A video containing several funny video snippets.

Happy-2.flv: A police playing jokes on passers-by.

Happy-3.flv: An American man playing jokes on passers-by with paper money attached to his foot. He pretends to be in a deep sleep to test who will take away the money.

Happy-4.flv: An old woman playing tricks on a man.

Table 6 The number of users of different playlists under three illuminations

Playlist no.	Number of subjects		
	Frontal	Left	Right
01	23	0	0
02	2	0	0
03	53	0	0
04	131	0	0
05	1	0	0
11	0	56	1
12	0	33	1
13	0	56	1
14	0	70	0
21	0	0	13
22	0	0	5
23	0	0	2
24	0	2	143
31	5	4	3
32	0	1	0

Table 7 Content of the playlists

Playlist no.	Content					
	1	2	3	4	5	6
01	Happy-1.flv	Disgust-1.wmv	Fear-1.flv	Surprise-1.flv	Sad-1.avi	Anger-1.flv
02	Happy-2.flv	Disgust-2.flv	Fear-2.flv	Surprise-2.flv	Sad-1.avi	Anger-1.flv
03	Happy-1.flv	Disgust-3.flv	Fear-3.flv	Surprise-3.flv	Sad-2.avi	Anger-1.flv
04	Happy-1.flv	Disgust-3.flv	Fear-3.flv	Surprise-3.flv	Sad-2.avi	Anger-1.flv
05	Happy-4.flv	Disgust-4.avi	Fear-4.flv	Surprise-4.flv	Sad-3.flv	Anger-2.flv
11	Surprise-5.flv	Sad-4.flv	Fear-5.flv	-	-	-
12	Happy-3.flv	Disgust-5.flv	Fear-5.flv	Surprise-5.flv	Sad-4.flv	Anger-3.flv
13	Disgust-6.flv	Anger-4.flv	Happy-5.flv	-	-	-
14	Surprise-5.flv	Sad-4.flv	Fear-5.flv	Disgust-7.avi	Anger-3.flv	Happy-3.flv
21	Anger-5.flv	Sad-5.flv	Disgust-8.flv	Surprise-6.flv	Fear-6.flv	Happy-6.flv
22	Sad-5.flv	Anger-5.flv	Disgust-8.flv	Surprise-6.flv	Fear-6.flv	Happy-6.flv
23	Sad-5.flv	Anger-5.flv	Disgust-8.flv	Surprise-2.flv	Fear-6.flv	Happy-6.flv
24	Sad-5.flv	Anger-5.flv	Disgust-8.flv	Surprise-2.flv	Fear-6.flv	Happy-6.flv
31	Happy-2.avi	Disgust-1.wmv	Fear-1.flv	Surprise-7.avi	Sad-6.flv	Anger-4.flv
32	Happy-7.flv	Disgust-9.flv	Fear-7.flv	Surprise-6.flv	Sad-7.flv	Anger-4.flv

Happy-5.flv: A news vendor playing tricks on passers-by by hiding his head when people come to ask for help. Happy-6.flv: An American playing tricks on passers-by. He puts glue on the chair and waits for people to sit. When the people stand up, their pants are torn.

Happy-7.flv: Two Americans playing tricks on a fitness instructor at a fitness club. They put one mirror in front of a wall. When someone shows his or her figure in front of the mirror, they slowly push the mirror down toward the person.

Disgust-1.wmv: A video showing the process of creating a crocodile tattoo in Africa, which contains some scenes that may induce a feeling of disgust.

Disgust-2.flv: A bloody cartoon movie containing some unsettling scenes.

Disgust-3.flv: Nauseating film snippets containing some disturbing scenes.

Disgust-4: A bloody cartoon movie that may cause discomfort.

Disgust-5.flv: A disturbing video showing a man take his heart out.

Disgust-6.flv: A cartoon movie, Happy Tree Friends, containing many bloody and disgusting scenes.

Disgust-7.avi: A puppet show containing some bloody and disgusting scenes.

Disgust-8.flv: A video showing a man eating a large worm.

Disgust-9.flv: A bloody cartoon movie that may cause discomfort.

Fear-1.flv: A daughter scaring her father with a dreadful face.

Fear-2.flv: A short video relating a ghost story about visiting a friend.

Fear-3.flv: A video of a dreadful head appearing suddenly after two scenery images are displayed.

Fear-4.flv: A video of a dreadful head appearing suddenly out of a calm scene.

Fear-5.flv: A short video relating a ghost story that takes place in an elevator.

Fear-6.flv: A short video relating a ghost story that takes place when visiting a friend.

Fear-7.flv: A video of a dreadful head appearing suddenly in a messy room.

Surprise-1.flv: A Chinese magician performing a surprising magic trick: passing through a wall without a door.
Surprise-2.flv: A magician removing food from a picture of ads on the wall.
Surprise-3.flv: A video of a magic performed on America's Got Talent: a man is sawed with a chainsaw.
Surprise-4.flv: A collection of amazing video snippets.
Surprise-5.flv: A video clip showing amazing stunts segmented from a TV show.
Surprise-6.flv: A video of a man creating a world in an inconceivable way; the video appears to be a clip from a science-fiction film.
Surprise-7.avi: A video showing an amazing motorcycle performance.
Sad-1.avi: A video showing pictures of the China 512 Earthquake.
Sad-2.avi: A video showing sad pictures of the China 512 Earthquake.
Sad-3.flv: A video showing some heart-warming video snippets of the China 512 Earthquake.
Sad-4.flv: A video showing 100 sad scenes of the China 512 Earthquake.
Sad-5.flv: A video relating the facts of the Japanese invasion of China during the Second World War.
Sad-6.flv: A video showing touching words spoken by children when the China 512 Earthquake occurred.
Sad-7.flv: A video showing touching words spoken by Wen Jiabao, premier of China, when the China 512 Earthquake occurred.
Anger-1.flv: A video of a brutal man killing his dog.
Anger-2.flv: A video of students bullying their old geography teacher.
Anger-3.flv: A video showing a disobedient son beating and scolding his mother in the street.
Anger-4.flv: A video showing a Japanese massacre in Nanjing during the Second World War.
Anger-5.flv: A video cut from the film The Tokyo Trial, when Hideki Tojo is on trial.

Table 8 The subject number in each illumination directory

Illumination	Subject number
Front	7 8 11 12 13 15 21 28 31 36 39 43 44 63 71 75 76 79 80 81 82 83 84 87 91 92 96 97 98 101 102 103 104 105 107 108 109 111 113 114 116 117 118 119 120 121 123 124 125 126 127 128 129 133 135 136 137 139 140 143 144 146 149 150 151 153 154 156 158 160 161 162 165 166 167 168 171 173 174 175 177 178 179 180 182 183 185 186 187 190 191 192 194 196 197 198 200 201 203 205 209 210
Left	4 10 13 14 23 25 30 31 36 39 41 42 44 64 69 71 74 75 76 82 83 85 87 89 90 92 96 97 102 103 104 105 107 108 109 113 114 116 117 118 119 123 124 125 126 127 128 129 133 135 139 140 143 144 146 149 150 151 154 156 158 159 160 161 162 165 171 175 177 178 179 180 182 183 185 186 187 188 190 191 192 194 196 197 200 201 203 204 205 209 210
Right	3 4 8 10 14 15 28 31 36 39 44 49 52 59 60 66 70 71 73 76 79 90 91 92 96 97 101 102 103 104 105 107 108 109 111 113 114 116 117 119 120 121 123 124 125 126 127 128 129 133 135 137 139 140 143 144 149 150 151 154 156 158 159 161 162 165 166 168 171 172 174 175 178 179 180 182 183 185 186 187 188 190 191 192 196 197 200 201 203 204 205 209 210

The subject numbers in each illumination directory in the released NVIE database are shown in Table 8.

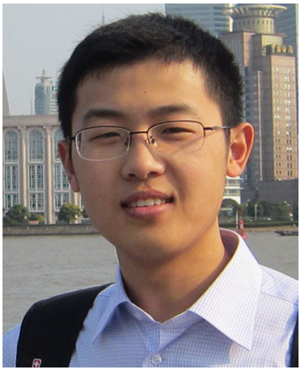
References

1. Arapakis I, Konstas I, Jose JM (2009) Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In: Proceedings of the 17th ACM international conference on multimedia, MM '09. ACM, New York, pp 461–470
2. Arapakis I, Moshfeghi Y, Joho H, Ren R, Hannah D, Jose JM (2009) Enriching user profiling with affective features for the improvement of a multimodal recommender system. In: Proceedings of the ACM international conference on image and video retrieval, CIVR '09. ACM, New York, pp 29:1–29:8
3. Arapakis I, Moshfeghi Y, Joho H, Ren R, Hannah D, Jose JM (2009) Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In: Proceedings of the 2009 IEEE international conference on multimedia and expo, ICME'09. IEEE Press, Piscataway, pp 1440–1443
4. Canini L, Gilroy S, Cavazza M, Leonardi R, Benini S (2010) Users' response to affective film content: a narrative perspective. In: 2010 international workshop on content-based multimedia indexing (CBMI). pp 1–6
5. Coan JA, Allen JJB (2007) Handbook of emotion elicitation and assessment. Oxford University Press, New York
6. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685
7. Hanjalic A, Xu L-Q (2005) Affective video content representation and modeling. IEEE Trans Multimed 7(1):143–154
8. Joho H, Jose JM, Valenti R, Sebe N (2009) Exploiting facial expressions for affective video summarisation. In: Proceedings of the ACM international conference on image and video retrieval, CIVR '09. ACM, New York, pp 31:1–31:8
9. Kierkels JJM, Soleymani M, Pun T (2009) Queries and tags in affect-based multimedia retrieval. In: Proceedings of the 2009 IEEE international conference on multimedia and expo, ICME'09. IEEE Press, Piscataway, pp 1436–1439
10. Koelstra S, Muehl C, Patras I (2009) EEG analysis for implicit tagging of video data. In: Workshop on affective brain-computer interfaces. Proceedings ACII, pp 27–32
11. Koelstra S, Muhl C, Soleymani M, Lee J-S, Yazdani A, Ebrahimiv T, Pun T, Nijholt A, Patras I (2012) Deap: a database for emotion analysis: using physiological signals. IEEE Trans Affect Comput 3:18–31
12. Koelstra S, Yazdani A, Soleymani M, Mühl C, Lee J-S, Nijholt A, Pun T, Ebrahimi T, Patras I (2010) Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In: Proceedings of the 2010 international conference on brain informatics, BI'10. Springer, Berlin, pp 89–100
13. Kreibig SD (2010) Autonomic nervous system activity in emotion: a review. Biol Psychol 84(3):394–421
14. Krzywicki AT, He G, O'Kane BL (2009) Analysis of facial thermal variations in response to emotion-eliciting film clips. In: Quantum information and computation VII, the international society for optical engineering. SPIE, pp 734312–734312–11
15. Liu Z, Wang S, Wang Z, Ji Q (2013) Implicit video multi-emotion tagging by exploiting multi-expression relations. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp 1–6
16. Lv Y, Wang S, Shen P (2011) A real-time attitude recognition by eye-tracking. In: Proceedings of the third international conference on internet multimedia computing and service, ICIMCS '11. ACM, New York, pp 170–173

17. Money AG, Agius H (2009) Analysing user physiological responses for affective video summarisation. Displays 30(2):59–70. cited By (since 1996) 8
18. Money AG, Agius HW (2008) Feasibility of personalized affective video summaries. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction, volume 4868 of Lecture Notes in Computer Science. Springer, pp 194–208
19. Money AG, Agius HW (2010) Elvis: entertainment-led video summaries. ACM Trans Multimed Comput Commun Appl 6(3):17:1–17:30
20. Ong K-M, Kameyama W (2009) Classification of video shots based on human affect. Inf Media Technol 4(4):903–912
21. Pantic M, Vinciarelli A (2009) Implicit human-centered tagging. IEEE Signal Proc Mag 26(6):173–180
22. Peng W-T, Chang C-H, Chu W-T, Huang W-J, Chou C-N, Chang W-Y, Hung Y-P (2010) A real-time user interest meter and its applications in home video summarizing. In: 2010 IEEE international conference on multimedia and expo (ICME), pp 849–854
23. Peng W-T, Chu W-T, Chang C-H, Chou C-N, Huang W-J, Chang W-Y, Hung Y-P (2011) Editing by viewing: automatic home video summarization by viewing behavior analysis. IEEE Trans Multimed 13(3):539–550
24. Peng W-T, Huang W-J, Chu W-T, Chou C-N, Chang W-Y, Chang C-H, Hung Y-P (2008) A user experience model for home video summarization. In: Proceedings of the 15th international multimedia modeling conference on advances in multimedia modeling, MMM '09. Springer, Berlin, pp 484–495
25. Rainville P, Bechara A, Naqvi N, Damasio AR (2006) Basic emotions are associated with distinct patterns of cardiorespiratory activity. Int J Psychophysiol 61(1):5–18
26. Scott I, Cootes T, Taylor C. Aam modelling and search software. http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools_doc/index.html
27. Smeaton AF, Rothwell S (2009) Biometric responses to music-rich segments in films: the cdvplex. In: International workshop on content-based multimedia indexing, pp 162–168
28. Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence, pp 1015–1021
29. Soleymani M, Chanel G, Kierkels J, Pun T (2008) Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Tenth IEEE international symposium on multimedia, 2008. ISM 2008, pp 228–235
30. Soleymani M, Chanel G, Kierkels JJM, Pun T (2008) Affective ranking of movie scenes using physiological signals and content analysis. In: Proceedings of the 2nd ACM workshop on multimedia semantics. MS '08. ACM, New York, pp 32–39
31. Soleymani M, Koelstra S, Patras I, Pun T (2011) Continuous emotion detection in response to music videos. In: 2011 IEEE international conference on automatic face gesture recognition and workshops (FG 2011), pp 803–808
32. Soleymani M, Lichtenauer J, Pun T, Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3(1):42–55
33. Toyosawa S, Kawai T (2010) An experience oriented video digesting method using heart activity and its applicable video types. In: Proceedings of the 11th Pacific Rim conference on advances in multimedia information processing: part I. PCM'10. Springer, Berlin, pp 260–271
34. Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, Chen F, Wang X (2010) A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimed 12(7):682–691
35. Wang S, Wang X (2010) Kansei engineering and soft computing: theory and practice, chapter 7: emotional semantic detection from multimedia: a brief overview. IGI Global, Pennsylvania
36. Yazdani A, Lee J-S, Ebrahimi T (2009) Implicit emotional tagging of multimedia using EEG signals and brain computer interface. In: Proceedings of the first SIGMM workshop on social media. WSM '09. ACM, New York, pp 81–88
37. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H (2010) Advancing feature selection research–asu feature selection repository. Technical report, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe

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.

Zhilei Liu, received the Bachelor degree in School of Mathematics and Information Science from the Shandong University of Technology, Zibo, Shandong Province, China, in 2008. And he is studying for Ph.D degree in School of Computer Science and Technology from the University of Science and Technology of China (USTC), Hefei, Anhui Province, China. His research interest is Affective Computing.



Yachen Zhu, received the Bachelor degree in School of Computer Science and Technology from University of Science and Technology of China, Hefei, Anhui Province, China, in 2010. And he continues studying for PhD degree there. His research interest is Affective Computing.

Menghua He, received the Bachelor degree in School of Mathematical Sciences from Anhui University, Hefei, China, in 2011. And she is studying for Master degree in School of Computer Science and Technology from the University of Science and Technology of China (USTC), Hefei, Anhui Province, China. Her research interest is Affective Computing.

Xiaoping Chen, received a PhD in Computer Science from USTC, a M.E. in Electrical Engineering and a B.A. in Mathematics from Anhui University. Prof. Chen is currently Director of the Center for Artificial Intelligence Research, USTC. He also serves as Trustee of the International RoboCup Federation, Member of Editorial Board of Journal of Artificial Intelligence Research, Member of Editorial Board of Knowledge Engineering Review, and Chair of the Chinese National RoboCup Committee. Prof. Chen has been focused on the research and teaching in the field of Artificial Intelligence and Autonomous Robotics. He established and has led the USTC Multi-Agent Systems Lab and robot team, WrightEagle. With these platforms, Prof. Chen found the KeJia Project in 2008, which aims at developing a human-level intelligent service robot, in a sense similar to what Alan Turing specified in his “Thinking Machine”.

Qiang Ji, received his PhD 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 Deparment 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 are in computer vision, probabilistic graphical models, pattern recognition, information fusion for situation awareness and decision making under uncertainty, human computer interaction, and robotics.