Multimed Tools Appl
DOI 10.1007/s11042-013-1450-8

Hybrid video emotional tagging using users' EEG and video content

Shangfei Wang · Yachen Zhu · Guobing Wu · Qiang Ji

Abstract In this paper, we propose novel hybrid approaches to annotate videos in valence and arousal spaces by using users' electroencephalogram (EEG) signals and video content. Firstly, several audio and visual features are extracted from video clips and five frequency features are extracted from each channel of the EEG signals. Secondly, statistical analyses are conducted to explore the relationships among emotional tags, EEG and video features. Thirdly, three Bayesian Networks are constructed to annotate videos by combining the video and EEG features at independent feature-level fusion, decision-level fusion and dependent feature-level fusion. In order to evaluate the effectiveness of our approaches, we designed and conducted the psychophysiological experiment to collect data, including emotion-induced video clips, users' EEG responses while watching the selected video clips, and emotional video tags collected through participants' self-report after watching each clip. The experimental results show that the proposed fusion methods outperform the conventional emotional tagging methods that use either video or EEG features alone in both valence and arousal spaces. Moreover, we can narrow down the semantic gap between the low-level video features and the users' high-level emotional tags with the help of EEG features.

S. Wang (✉) · Y. Zhu · G. Wu
School of Computer Science and Technology, University of Science and Technology of China,
Hefei, Anhui, China
e-mail: sfwang@ustc.edu.cn

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

G. Wu
e-mail: guobing@mail.ustc.edu.cn

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

Published online: 10 April 2013

Keywords Emotional tagging · Videos · Independent feature-level fusion · Decision-level fusion · Dependent feature-level fusion

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 classification and retrieval of digital videos, assigning them emotional tags has been an active research area in recent decades [80]. Video tagging is usually divided into two categories: explicit and implicit tagging [76]. 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 [76].

Since the research of emotional tagging of videos deals with videos and human's emotional response while watching the videos, it should consist of video content space, users' spontaneous response space, and emotional tag space, as well as their relationships [80]. Present explicit tagging approach involves the mapping between video content space and emotional tag space, while the emerging implicit tagging approach considers the mapping between users' spontaneous response space and emotional tag space. We believe that fully exploiting the three spaces and their relationships is crucial to reducing the semantic gap between the low-level video features and the users' high-level emotional tags.

In this paper, we propose novel emotional tagging approaches combining video content and physiological signals in three ways (i.e., independent feature-level fusion, decision-level fusion, and dependent feature-level fusion). Firstly, several EEG features and video features are extracted. Secondly, statistical analyses are conducted to explore the relationships among emotional tags and video and EEG features. After that, three kinds of Bayesian Networks (BN) are constructed for emotional tagging. The first one is an independent feature-level fusion, in which EEG and video features are independent given emotional tags. The second performs decision-level fusion by combining the tagging results from EEG signals and video features. The third performs dependent feature-level fusion by directly considering the relationship between EEG and video features. To evaluate our approaches, we designed and conducted a psychophysiological experiment to collect data. An emotion-induced video dataset is gathered, which consists of 92 video clips with intrinsic emotional content from movies and TV series. Subjects' EEG signals are recorded while they watch video clips. Ground truth emotional tags of videos in terms of valence and arousal are collected by asking subjects to report their induced emotions immediately after watching every clip. The experimental results demonstrate that our proposed fusion methods outperform the conventional emotional tagging methods that use either video or EEG features alone in both valence and arousal spaces. Independent feature-level fusion yields the best performance. The comparison of the emotional tagging method using video alone and the dependent feature-level fusion method confirms the role of narrowing semantic gaps by adopting users' physiological signals.

The outline of the paper is as follows. In Section 2, we briefly review the related works on video emotional tagging and multi-modality fusion for emotion recognition. In addition, we propose a framework for emotional tagging of videos.

Section 3 presents the construction of three spaces and analyses of their relationships. Section 4 elaborates on the proposed emotion tagging methods. Section 5 discusses the experimental results and analyses. Section 6 concludes the paper.

2 Related work

2.1 Video emotional tagging

To the best of our knowledge, study on emotional tagging of videos is first conducted at the beginning of the last decade by Moncrieff et al. [54]. Since then, hundreds of papers have been published in this field. Earlier studies mainly infer emotional tags from video content, and so use the explicit approach.

Two kinds of tag descriptors are often used. One is the categorical approach [27, 32, 33, 41, 54, 60, 72, 78, 81–83, 89]. It uses the six basic emotions (happiness, sadness, surprise, fear, disgust and anger) as well as adjectives and adjective pairs, such as pleasing, boring, or irritating. The other is the dimensional approach [4–6, 23, 73, 79, 85, 90, 91], such as valence and arousal [19, 24, 69].

Various kinds of visual and audio features (e.g., color, motion, and sound energy) are extracted from videos [54, 73, 82]. The mapping from video features to emotional tags are modeled by different machine learning methods, such as support vector machine [82], support vector regression [15], neural networks [81], hidden Markov model [72, 85, 86], dynamic Bayesian networks [5], conditional random fields [87], etc.

Only recently have researchers begun to realize that users' spontaneous physiological and behavior responses are useful hints to videos' emotional tags. By recognizing users' emotion from their induced physiological/behavioral signals while watching the videos, the tags can be obtained automatically. This is called the implicit approach. Two pioneer groups, Money et al. [55, 56] and Soleymani et al. [68, 70], have investigated many kinds of physiological signals as the implicit feedback, including electroencephalography, electromyography, galvanic skin resistance, blood pressure, respiration, heart rate, and skin temperature, while some other researchers only focus on one or two kinds of physiological signals [14, 64, 67]. Two other groups consider event-related potential (ERP), such as N400 [42] and P300 [88], as subjects' implicit feedback. Several researchers also consider implicit tagging according to spontaneous visual behavior, such as facial expressions and eye gaze [2, 3, 30, 46, 61, 62]. A comprehensive overview can be found in [50, 80].

Although researchers believe emotional tagging of videos involves videos' content, users' spontaneous responses, and users' subjective evaluations (that is videos' emotional tags), few researchers have fully explored the relationships between the three. Soleymani [70] analyzed the relationship between subjects' physiological response and videos' emotional tag in both arousal and valence, as well as the relationship between the content of videos and the emotional tag in arousal and valence. In this paper, we propose a framework of emotional tagging of videos that fully exploits these three spaces and their relationships as shown in Fig. 1. The physical space includes various visual and audio features extracted from videos to represent content; the psychological space captures the users' subjective tagging of the videos' emotional content in terms of dimension or category (this may be also called emotional tag space); and the physiological space represents users'

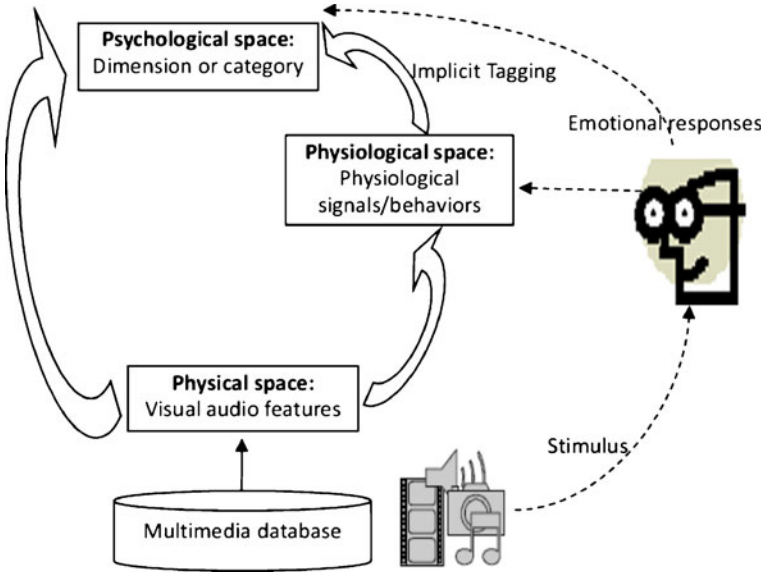


Fig. 1 General framework for emotional video tagging [80]

physiological and behavioral responses while watching the videos. The mapping from physical space to psychological space can be regarded as explicit tagging, while the mapping from physiological space to psychological space is implicit tagging. The relationship between physical space and physiological space has rarely been studied, although it may help reduce the semantic gap between low-level video features and the high-level users' high-level emotional tags, thus improving the performance of emotional tagging.

2.2 Multimodality fusion for emotion recognition

As we know, humans interpret others' emotional states using multiple modalities such as facial expression, speech, behavior, and posture. Thus, many researchers believe that computers should also recognize users' emotions by fusing multiple modalities.

Generally speaking, multimodal data fusion can be accomplished at three levels: data, feature, and decision. Data-level fusion directly integrates raw data of multiple modalities. It requires data of the same type. Feature-level fusion is performed on the features extracted from each modality. It is appropriate for closely coupled and synchronized modalities. The features extracted should be synchronous and compatible. Decision-level fusion is based on the fusion of the individual mode decisions. It is considered to be the most robust and resistant to individual sensor and is computationally less expensive than feature fusion.

Current works on multimodal emotion recognition have adopted both feature-level [11, 17, 18, 21, 40] and decision-level fusion [7, 11, 18, 31, 34–36, 40].

The combined modality can be face-body [34], face-physiological signal [8], face-speech [51], speech- physiological signal [39], multiple channel physiological signal [1, 12, 22, 25, 48, 49, 59, 75, 77], face- voice-body [10] and speech-text [20]. Many machine learning methods are adopted, such as support vector machine (SVM) [7, 20, 34, 35, 82], neural network [17, 36, 81], linear discriminant analyses [21, 40], Gaussian process classification [7, 34, 35], Bayesian classifier [18], hidden Markov models [33, 47, 72, 85, 86], and dynamic Bayesian networks [5]. Most reported work achieved better performance with feature-level fusion [26]. A comprehensive overview can be found in [13, 66, 92].

For explicit emotional video tagging, most current works have adopted feature-level fusion by concatenating audio and visual features as the input of a classifier [23]. A few works employ decision-level fusion by combining the tagging results from visual and audio features to yield final fused tagging results [57, 73, 86].

For implicit emotional video tagging, if multiple modalities of users' physiological and behavioral responses are adopted to infer the video tag, the data- fusion approaches are similar as those used in multimodal emotion recognition.

Until now, there has been no reported research integrating video content and users' EEG signals to annotate videos. Compared to previous work, our contributions are summarized as follows: (1) It is the first instance of combining video content and users' physiological response at two different levels (feature-level and decision-level) to annotate videos' emotional tags. (2) We investigate and model relationships between video and EEG features, which can narrow down the semantic gap between the low-level physical features and the users' high-level emotional tags.

3 Emotional video analyses

Figure 2 shows the framework of how to construct three spaces and analyze their relationships. After the psychophysiological experiment and data collection, EEG features and video features are extracted. Then, statistical analyses for hypothesis testing are conducted to check whether there are significant differences in every feature between the two groups of emotional tags. Correlation analysis between EEG and video features is also employed.

3.1 Emotional tag space

The descriptors of emotions are valence and arousal. Valence refers to how positive or negative an event is, and arousal reflects whether an event is exciting/agitating or calming/soothing [38]. In this work, valence and arousal are divided into two categories: positive/negative valence, and high/low arousal.

3.2 Physiological space

The physiological space consists of several EEG features. First, noise mitigation is carried out. The Horizontal and Vertical Electro-OculoGram (HEOG and VEOG) are removed and a band pass filter with a lower cutoff frequency of 0.3 Hz and a

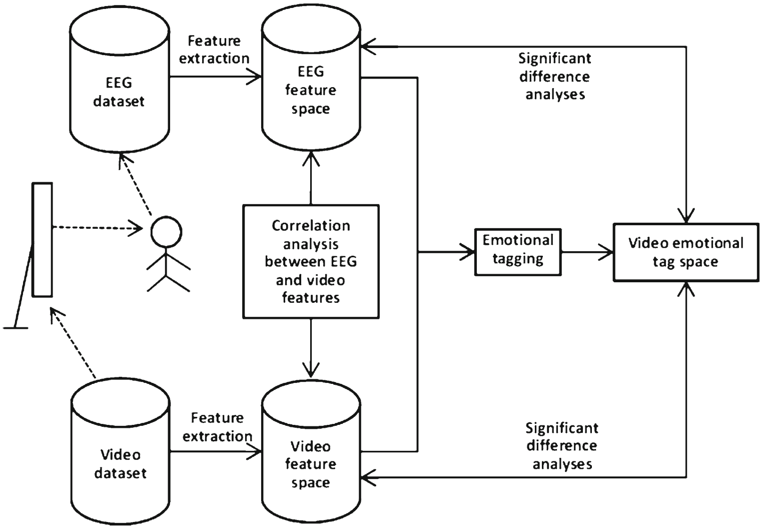

Feature extraction
EEG dataset
EEG feature space
Significant difference analyses
Correlation analysis between EEG and video features
Emotional tagging
Video emotional tag space
Feature extraction
Video dataset
Video feature space
Significant difference analyses

Fig. 2 Framework for constructing and analyzing the three spaces

higher cutoff frequency of 45 Hz is used to remove DC drifts and suppress the 50Hz power line interference [42, 43]. Then, the power spectrum (PS) is calculated and divided into five segments [29]: the delta (0–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz) and gamma (30–45 Hz) frequency bands. The ratios of the power in each frequency band to the overall power are extracted as the features.

3.3 Physical space

Physical space consists of several visual and audio features that represent video content. For visual features, lighting, color, and motion are powerful tools to establish the mood of a scene and affect the emotions of the viewer according to cinematography and psychology. Thus, three features, named lighting key, color energy, and visual excitement are extracted from video clips [78]. Thirty-one commonly used audio features are extracted from the video, including average energy, average energy intensity, spectrum flux, Zero Crossing Rate (ZCR), standard deviation of ZCR, twelve Mel-frequency Cepstral Coefficients (MFCCs), log energy of MFCC, and the standard deviations of the thirteen MFCCs [53].

3.3.1 Visual features

Visual excitement represents the average number of pixels changed between corresponding frames according to human perception. The change is computed in the perceptually nearly uniform CIE Luv space [52] according to (1).

$$\frac{10}{N_c} * \sum_{f=1}^{N_c} \left[\sum_{k=1}^{N_H} \frac{H(x_{fd}(k) - thres_{fd})}{N_H} + \sum_{k=1}^{N_H} \frac{H(x_{fd}(k) - thres_{fd})}{N_H}^{W} \right]_f \qquad (1)$$

$$x_{fd} = \sqrt{s_L(L_1 - L_0)^2 + \frac{1}{3[(u_1 - u_0)^2 + (v_1 - v_0)^2]}}$$

$$s_L = \begin{cases} \frac{1}{3} & s_{avL} \geq \frac{1}{3} \\ \frac{1}{3} + \left(s_{avL} - \frac{1}{3}\right)^2 & \text{otherwise} \end{cases}$$

where (L_1, u_1, v_1) and (L_0, u_0, v_0) are the average CIE Luv values of corresponding blocks of consecutive frames, $thres_{fd}$ is the threshold, s_{avL} is the average frame luminance, N_c is the number of frames and k is the index of the frame, W is a constant, N_H is the block number of each frame and H is the Heaviside step function.

The lighting key is the product of the mean and variance of the brightness of a frame [63]. Suppose μ_i and σ_i are the mean and standard deviation of the brightness of the HSV space [16] respectively. The lighting key is expressed as

$$\xi_i = \mu_i \cdot \sigma_i \tag{2}$$

The mean value for the lighting key of all frames in a clip is calculated as the feature.

Color energy is defined as the product of the raw energy and color contrast:

$$\sum_i \sum_j p(c_i) \times p(c_j) \times d(c_i, c_j) \times \sum_k^M E(h_k) s_k v_k, \tag{3}$$

where c is a histogram bin indexed by i, j to iterate over every single bin index in HLS histogram of an image. $p(\cdot)$ is the histogram probability. $d(c_i, c_j)$ is the L 2-norm in HLS space while M is the total number of pixels, over which index k iterates. s_k, v_k are the corresponding saturation and lightness values while $E(h_k)$ is the energy of the hue depending on its angular distance to blue and red.

The detailed explanation and parameter setting of these three visual features can be found in [78].

3.3.2 Audio features

Suppose $x(t), t \subseteq [t_1, t_2]$ is an audio clip separated from a video clip. Firstly, we divide the audio clip into several segments using the sliding window with length of T and 50 % overlap. Secondly, we obtain the Fourier coefficients through Fast Fourier Transform Algorithm (FFT). Let $x_t(n), n \subseteq [0, N]$ be the time sequence data of the t-th segments, and $N_t(n)$ be the corresponding Fourier coefficient. Then, we calculate average energy, average energy intensity, spectrum flux, and zero crossing rate according through (4) to (7) respectively.

Average energy:

$$\int_{t1}^{t2} x^2(t)dt, \tag{4}$$

Average energy intensity: In physical perspective, the sound wave spreads in a spherical space, so the energy depends on where we are in relation to the source. If the distance is R, then the intensity is

$$\frac{1}{4\pi R^2}\int x^2(t)dt,$$

However, this is not the model for humans' auditory sense. The intensity in air, expressed in dB should be relative to the auditory threshold.

$$10\log_{10}\left(\frac{1}{P_0^2}\int x^2(t)dt\right), \tag{5}$$

where $x(t)$ is the sound pressure in units of Pa (Pascal), and $P_0 = 2\cdot 10^{-5}$Pa is the auditory threshold pressure.

Spectrum flux:

$$F_t = \sum_{n=1}^{N}(N_t(n) - N_{t-1}(n))^2, \tag{6}$$

Zero crossing rate:

$$Z_t = \frac{1}{2N}\sum_{n=1}^{N}|sign(x_t(n)) - sign(x_t(n-1))|, |x_t(n) - x_t(n-1)| > \delta, \tag{7}$$

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where δ is a threshold to check whether two adjacent frames have significant difference. We also calculate the standard deviation of ZCR.

In order to extract MFCCs, we first map the powers of the spectrum obtained from FFT onto the Mel scale, using triangular overlapping windows. We then take the logs of the powers at each of the Mel frequencies. From this, we take the discrete cosine transform of the list of Mel log powers. Thus, the first twelve amplitudes of the resulting spectrum are obtained. After that, we also calculate the log energy of MFCC, and the means and standard deviations of MFCCs.

3.4 Relations between physical/physiological space and emotional tags

After feature extraction, we conduct statistical analyses for hypothesis testing to analyze whether there are significant feature differences between the two groups of emotional tags. The null hypothesis H0 means the median difference between positive and negative valence (or high and low arousal) for a feature is zero. The alternative hypothesis H1 is that the median difference between positive and negative valence (or high and low arousal) for a feature is not zero. We may reject the null hypothesis when the P-value is less than the significant level. The procedures are described as follows: First, a normality test is performed on each feature. If the feature is not normally distributed, a Kolmogorov-Smirnov test is used. Otherwise, homogeneity of the features is tested. If the variance is homogeneous, a T-test with homogeneity of variance is performed; otherwise, a T-test with inhomogeneity of variance is performed. In our study, the P-value threshold is set to 0.01. This procedure can also be used as feature selection.

3.5 Relations between physical space and physiological space

The Pearson correlation coefficient with its P-value is calculated between each EEG and video feature with a significant difference between two emotional tag groups. The P-value is the probability that the correlation coefficient is zero. Here, P-value threshold is also set to 0.01.

4 Video emotional tagging with Bayesian networks

Three BNs are constructed to annotate videos with emotional tags from three aspects: independent feature-level fusion, decision-level fusion and dependent feature-level fusion. As a probabilistic graphical model, BN can effectively capture the uncertainties in data and allows data from different modalities to be systematically represented and integrated. Furthermore, BN can easily realize dependent and independent feature-level fusion, while other methods just concatenate features of multiple modalities to realize feature-level fusion.

4.1 Independent feature-level fusion

We use a three-node BN to perform independent feature-level fusion. The structure is shown in Fig. 3a.

The model consists of two layers: the emotional tag layer and the feature layer. The node in the emotional layer is discrete with two states, representing positive and negative valence, or high and low arousal. The two nodes in the feature layers are continuous, representing EEG features (F_e) and video features (F_v) respectively. The relationship between videos' emotional tags and EEG/video features are established through links. Two features are assumed to be independent given video tags.

In the training step, the tag prior probability $P(C)$ and the likelihood $p(F_e|C = k)$ (the probability of the training sample F_e given their class C) and $p(F_v|C = k)$ are estimated from the training data; In the testing step, the posterior probability $P(C|F_e, F_v)$ is computed for each class C, and the class is recognized as the one with the highest posterior probability, i.e.,

$$\begin{aligned} C^* &= \arg\max_C P(C|F_e, F_v) \\ &= \arg\max_C P(C)P(F_e|C)P(F_v|C). \end{aligned} \tag{8}$$

4.2 Decision-level fusion

The structure of BNs used for decision-level fusion is shown in Fig. 3b. This model consists of three layers: the top emotional tag layer, the intermediate emotional tag

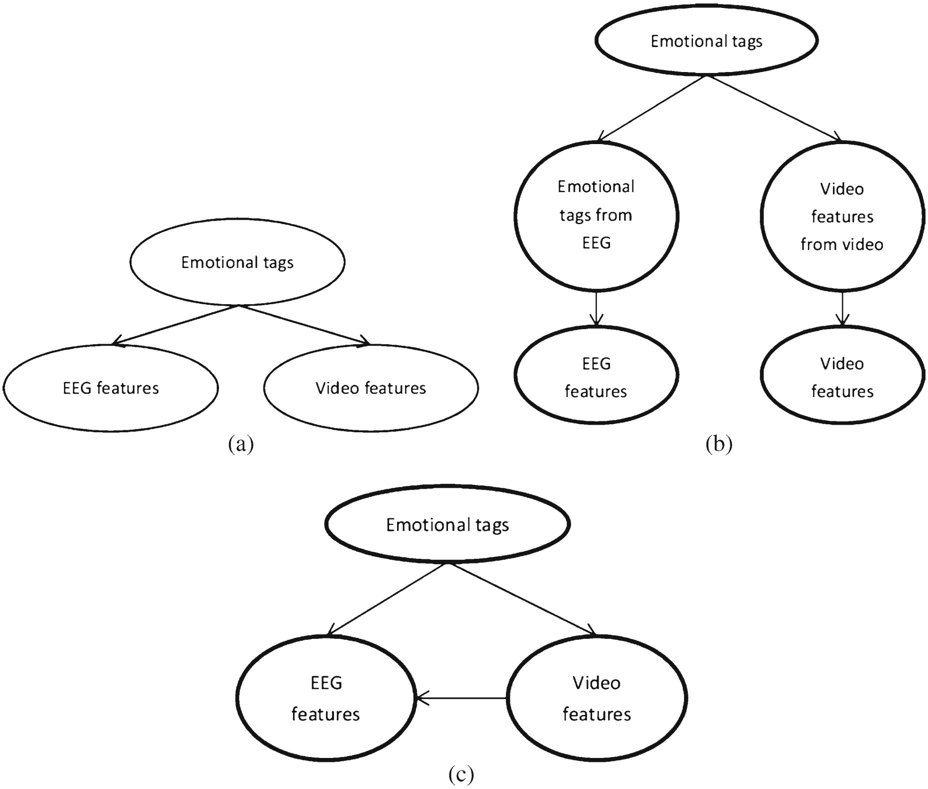


Fig. 3 **a** BN for independent feature-level fusion; **b** BN for decision-level fusion; **c** BN for dependent feature-level fusion

layer, and the feature layer. All nodes in the top layer and the second layer are discrete nodes with two states, and each state corresponds to the recognized emotional tags, i.e., positive and negative valence, or high and low arousal. The relations between final emotional tags, tags inferred from EEG signals, and tags inferred from video features are established through links, which capture the uncertainty of emotional tagging from EEG signals and video features. The lowest layer in the model is the feature layer containing EEG and video features. All variables in this layer are observable.

Given the above model structure, the conditional probability distribution (CPD) associated with each node in the model needs to be learned from the training set. Owing to lack of value of the second layer in the training set, we divided the parameter's learning phase into two phases.

In the first phase, the emotional tags are taken as the value of nodes in the intermediate emotional tag layer to learn the CPDs of nodes in the feature layer. The CPDs of the feature layer are parameterized as the multivariate Gaussian. Specifically, for the EEG features node, let F_e denote the EEG features and C_e denote the value of the EEG emotions node, and assume the CPD of EEG features $p(F_e|C_e = k)$ satisfying a multivariate Gaussian distribution with corresponding mean vector μ_k and covariance matrix $\sum_k$. It can be represented as [58]:

$$p(F_e|C_e = k) = \frac{1}{(2\pi)^{n/2}|\sum_k|^{1/2}} exp\left\{-\frac{1}{2}(F_e - \mu_k)^T \sum^{-1}(F_e - \mu_k)\right\}. \tag{9}$$

Then we can learn parameters μ_k and $\sum_k$ given emotion $C_e = k$ by the maximum likelihood estimation.

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_{ik}, \tag{10}$$

$$\hat{\sum}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}(x_{ik} - \hat{\mu}_k)(x_{ik} - \hat{\mu}_k)^T, \tag{11}$$

where N_k is the number of samples of emotions k, and x_{ik} is the feature vector of i-th samples of emotion k. Afterwards, we take the features in the training set as the input to obtain the intermediate results according to the CPD of feature nodes.

In the second phase, the conditional probabilistic table (CPT) for each node in the intermediate emotional tag layer is calculated based on the analysis of real emotions and the previously obtained intermediate results. Specifically, for the EEG emotional tag node,

$$p(C_e = i|C = j) = N_{ji}/N_j, \tag{12}$$

where N_j is the number of samples of emotion j in the training set, and N_{ji} denote the number of samples of emotion j whose EEG intermediate result is i. After training, given the video and EEG features of a video, we want to find the video emotional tag by maximizing the posterior probability of the emotional tag node as:

$$\begin{aligned} C^* &= \arg\max_C P(C|F_e, F_v) \\ &= \arg\max_C \sum_{C_e}\sum_{C_v} P(C)P(F_e|C_e)P(C_e|C)P(F_v|C_v)P(C_v|C), \end{aligned} \tag{13}$$

where C_e and C_v denote the states of the emotional tags inferred from EEG signals and video content respectively. F_e and F_v are the feature vectors of EEG signals and video clips. Therefore, the true state of emotional tags can be inferred through probabilistic inference.

4.3 Dependent feature-level fusion

Video and EEG features are independent given emotional tags in the above two fusion methods. In this model, we add a link between video features and EEG features to capture the relationship between two modalities directly, as shown in Fig. 3c. The distributions of video and EEG feature nodes are multivariate Gaussian and linear Gaussian respectively.

In the training step, the tag prior probability $P(C)$ and the likelihood $p(F_e|F_v, C = k)$ and $p(F_v|C = k)$ are estimated from the training set data; In the testing step, the posterior probability $P(C|F_e, F_v)$ is computed for each class C, and the class is recognized as the one with the highest posterior probability according to (14).

$$\begin{aligned} C^* &= \arg\max_C P(C|F_e, F_v) \\ &= \arg\max_C P(C)P(F_e|C, F_v)P(F_v|C). \end{aligned} \tag{14}$$

5 Experiments and analyses

5.1 Psychophysiological experiments and data collection

Due to the copyright issues, most datasets used in current emotional video tagging are self-collected and are not available for other researchers. Schaefer et al. [65], Koelstra et al. [44] and Soleymani et al. [69] have created emotional video datasets for public research. Their datasets provide the recorded physiological signals of subjects as they watch videos, as well as their subjective evaluations. However, the video stimuli are not provided. Thus, we designed and conducted the psychophysiological experiment to collect the data.

5.1.1 Subjects

Twenty eight healthy students, including 25 males and 3 females, with ages ranging from 18 to 28 years old, attended our experiments. All of them had normal vision. In addition, an informed consent form was signed before the experiment and subjects received compensation.

5.1.2 Stimuli

Video clips were taken as stimuli to induce subjects' emotions in our experiments. These were extracted from a variety of genres, including horror, comedy, action, and drama, and were taken from DVDs and the Internet. Each clip contains a full emotional event as judged by the authors. Three raters evaluated the primary emotion of these clips in six categories: happiness, fear, sadness, disgust, surprise, and anger. The clip was included in the dataset if all three raters agreed on the main emotional category. The Source video, type, start-end time and resolution of each clip is provided in Table 3 in the Appendix. After that, we randomly selected clips with different emotional categories from the dataset to construct the playlist for subjects. The number of clips that the subjects watched varied, but the length of the playlists was almost the same.

5.1.3 Experimental setup

The diagram of the experiment is illustrated in Fig. 4. An isolated room that shielded physical noises and electromagnetic effects was used to ensure the accuracy of the collected data. A 19-inch LCD screen was placed on the table, about three feet from the sofa. The room on the left was for control and observation.

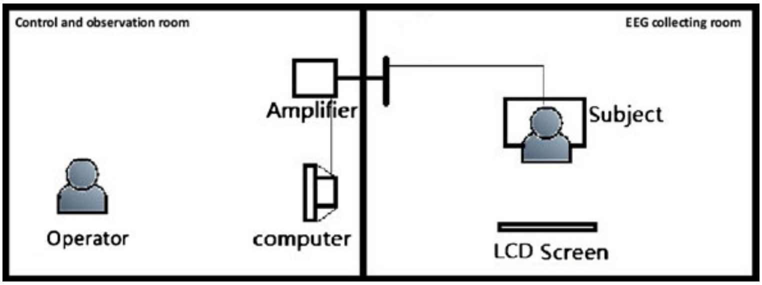


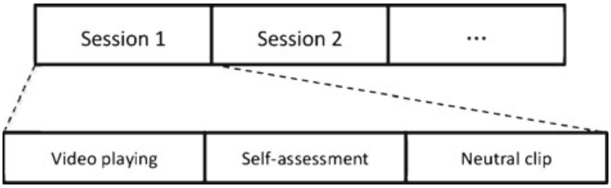
Fig. 4 Experimental environment for data collection

It contained two signal amplifiers and two computers utilized for recording the EEG signals and signal synchronization between EEG and video stimulus signal, respectively. EEG signals were acquired via a Quik-cap (Neuro Inc., El Paso, TX) with 32 Ag-AgCl electrodes arranged in an extended 10–20 system montage, which could record HEOG and VEOG, together with the usage of Neuroscan Synamps2 bio amplifiers and Neuroscan software (v4.3.1). The settings were as follows: sampling rate was 500Hz, amplifier gain was 1000, band-pass filter was between 0.3 and 100Hz without a notch filter, and mode was AC (Alternating Current) [42, 44, 71].

5.1.4 Experimental procedure

First, a detailed introduction was given to the subjects about the purpose of the experiment, the meaning and measurement scale of emotional arousal and valence, and the procedures of the experiment. The subject was asked to sit on a sofa in the isolated room seen in Fig. 4, and wore the Quik-cap and earphones. Experiments were performed according to the experimental arrangement illustrated in Fig. 5. The subject watched film clips from different emotional categories randomly selected from our video data set, and his/her EEG signals were collected synchronously. After watching each clip, the subject was asked to report his/her real induced emotion by filling out forms using emotional valence (−2, −1, 0, 1, 2) and arousal (−2, −1, 0, 1, 2). To reduce the interference between different emotions, video clips that would induce neither negative nor positive emotions, which was defined as neutral clips, were displayed to subjects during the intervals between two film clips of different categories.

Fig. 5 Experimental arrangement of stimuli



The video durations range from 34 s to 5 min 44 s. The subjects watched a different number of clips, but each spent the same amount of time watching video clips. However, a hardware problem in the EEG signal collecting device damaged some raw EEG signal segments, so the actual number of EEG segments for each subject ranges from one to five. Thus, we obtained 197 EEG segments and self-reported emotion recordings corresponding to 92 video clips.

5.2 Results and analyses

5.2.1 Details of sample distribution in valence-arousal space and EEG feature extraction

When a subject watched a video clip, we recorded his/her EEG signals and subjective evaluation. Thus, one EEG segment corresponds to one emotional tag in valence and arousal. Some videos were watched by multiple subjects, producing several subject-specific emotional tags. To produce one tag for each video, we averaged the several subject-specific emotion tags. Thus, there are 197 samples used for emotional tagging from EEG, 92 samples for emotional tagging from videos only, and 197 samples for emotional tagging by fusing EEG and videos.

We divided users' five-scale evaluations (−2, −1, 0, 1, 2) into two groups based on whether they are higher than zero or not. Thus, in arousal space, 149 EEG recordings are high, and 48 recordings are low; 70 videos are high and 22 are low. In valence space, 77 EEG recordings are positive, and 120 are negative; 30 video clips are positive and 62 are negative. Before extracting EEG features, four bad channels (T8, F3, FP1 and FP2) were removed from the original EEG data due to the hardware issue. The invalidated data of these four channels does not affect other channels.

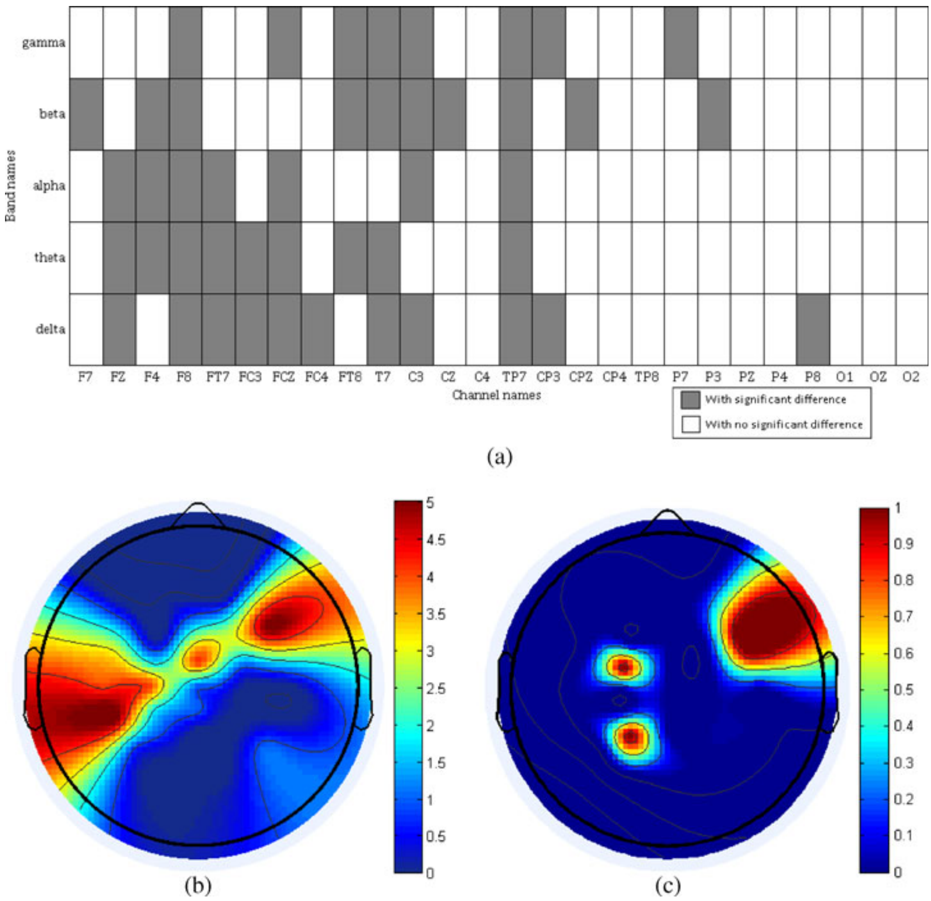
5.2.2 Relations between EEG features and emotional tags

The EEG features with significant differences between two valence groups are shown in Fig. 6a, where the gray cells represent the target features. The horizontal ordinate represents the electrodes and the vertical one represents the frequency bands used to extract EEG features. There are 45 features in total, including 11 delta, 9 theta, 7 alpha, 10 beta, and 8 gamma frequency band features. The distribution of these EEG features over the brain is shown in Fig. 6b.

In Fig. 6b, the colors range from blue to red, representing the features distributed in the brain areas from less present to more present. From Fig. 6b, we find that the features gather in the frontal and temporal areas, with some in the parietal area. This means that the frontal and temporal areas are highly correlated with human emotions [9, 37], and the parietal area is involved in emotion elicitation [28].

The EEG features with significant differences between two arousal groups are shown in Fig. 7a, and their distributions over the brain are shown in Fig. 7b.

There are 85 EEG features in total, including 18 features on delta frequency bands, 18 for theta, 20 for alpha, 15 for beta and 14 for gamma. The features mainly exist in the occipital and temporal areas, which means that these two areas are highly relevant to the excitement of human emotion [45].



Band names	F7	FZ	F4	F8	FT7	FC3	FCZ	FC4	FT8	T7	C3	CZ	C4	TP7	CP3	CPZ	CP4	TP8	P7	P3	PZ	P4	P8	O1	OZ	O2
gamma				■			■		■	■	■			■	■				■							
beta	■		■	■					■	■	■	■		■		■				■						
alpha		■	■	■	■		■				■			■												
theta		■	■	■	■	■	■		■	■				■												
delta		■		■	■	■	■	■		■	■			■	■								■			

Channel names

■ With significant difference

□ With no significant difference

(a)

(b)

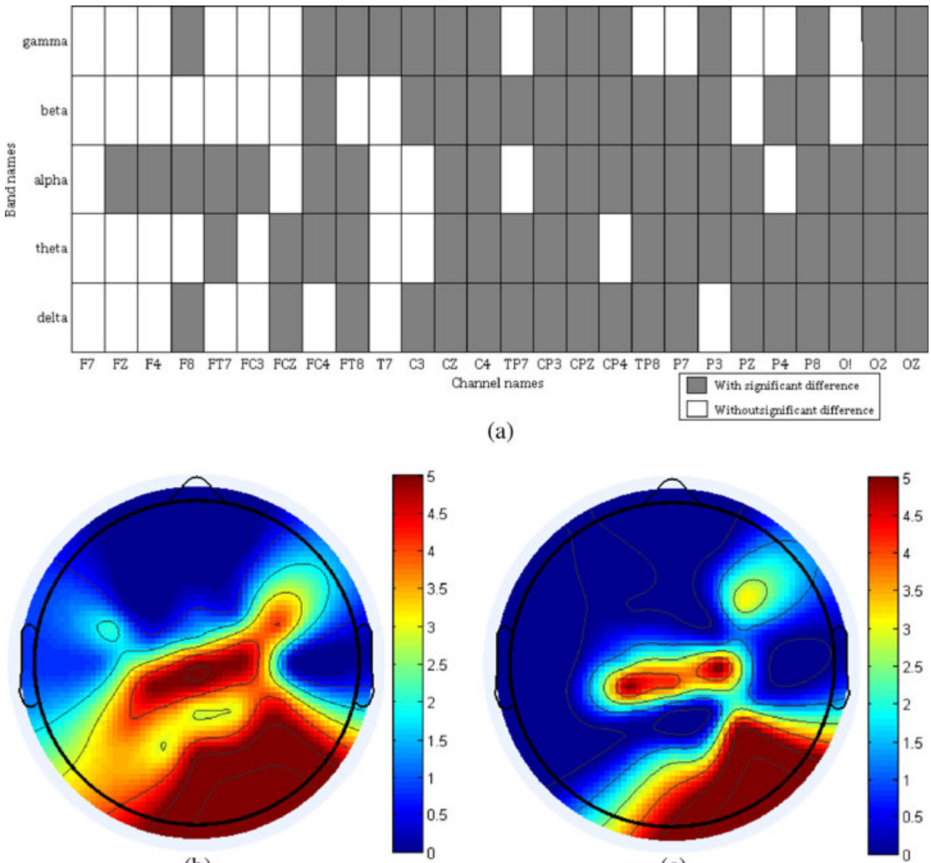
(c)

Fig. 6 **a** EEG features with significant differences in valence. **b** The distribution of EEG features with significant differences in valence. **c** The distribution of EEG features that have correlations in valance with video features

5.2.3 Relations between video features and emotional tags

The video features with significant differences between two valence groups are two visual features, including color energy and lighting key, and seven audio features, consisting of the 5th MFCC, the 8th, 9th, and 10th MFCCs standard deviation, the average energy, the average energy intensity, and the spectrum flux. For visual features, two of the three are selected, which means visual stimuli strongly influence emotions on valence [85]. For audio features, MFCC coefficients are discriminating features [85], since they occupy four out of seven among the selected audio features.

For arousal, the features with significant difference are nine audio features: the 2nd, 3rd, and 7th to 12th MFCCs, and the log energy of MFCC. This indicates that audio is closely connected to human emotions on arousal [78, 85], and further confirms MFCC features are important for emotional tagging [68]. No visual features are selected, possibly because only three visual features were measured.



(a)

(b)

(c)

Fig. 7 **a** EEG features with significant differences in arousal. **b** The distribution of EEG features with significant differences in arousal. **c** The distribution of EEG features that have correlations in arousal with video features

5.2.4 Relations between EEG and video features

Figures 6c and 7c show the distribution of EEG features, which are significantly correlated with video features on valence and arousal respectively. Thus, these features are more important for affect characterization. EEG features that correlate to video features for valence are listed as follows: (1) 8th MFCC's standard deviation correlates to P3's beta PS, FT8's gamma PS and F8's alpha PS; (2) spectrum flux correlates to F3's delta PS; (3) color energy correlates to FC3's delta PS.

In all, there are 57 EEG features with significant correlations to video features for arousal. All of the above shows that there is a close relationship between video content and EEG signals. For instance, the specific correlation coefficient between color energy and F3's delta PS is 0.1873, which means increasing red (energetic) or blue (relaxing) components results more in F3's delta band activity. This may be helpful for reducing the semantic gap between the low-level video features and the users' high-level emotional tags, and improving emotional tagging.

5.3 Emotional tagging experiments and results

To evaluate the effectiveness of our proposed tagging approaches, we compared the performance of three hybrid tagging methods with a tagging method using only EEG signals or video content.

To avoid overlap of samples in training and test phases, a leave-one-video-out cross-validation was used. Three parameters: accuracy, F1-score, and average precision, were adopted as the evaluation measurements.

5.3.1 Tagging results of valence

Tagging results of valence are shown in Table 1. From the table, we can see that:

(1) the accuracy and average precision of the tagging method using video content only is higher than those of the tagging method using EEG signals only. This is particularly true for positive samples, in which the accuracy using video is 73.33 % while the accuracy using EEG is 57.14 %. It seems that video content distinguishes positive and negative emotional tags better than EEG signals do.

(2) Both independent feature-level and decision-level fusion outperform tagging using only EEG signals or video content, since all three parameters of the two fusion methods are higher than those of tagging using one modality.

(3) Compared to emotion tagging from video content, dependent feature-level tagging that considers the relationship between video and EEG features performs better, since it has higher accuracy, F1-score, and average precision. This confirms that the emotional semantic gap can be narrowed by considering users' physiological signals.

Generally, our three hybrid tagging methods improve the performance of valence recognition.

Table 1 Tagging results of valence

	EEG only		Video only		Independent feature-fusion level	
	Positive	Negative	Positive	Negative	Positive	Negative
Positive	44	33	22	8	54	23
Negative	32	88	18	44	19	101
Accuracy	67.01 %		71.74 %		78.68 %	
F1-score	0.6524		0.6286		0.7200	
Average precision	0.5752		0.7215		0.7715	
	Decision-level fusion		Dependent feature-level fusion			
	Positive	Negative	Positive	Negative		
Positive	57	20	51	26		
Negative	24	96	20	100		
Accuracy	77.66 %		76.65 %			
F1-score	0.7215		0.6892			
Average precision	0.7701		0.7478			

Table 2 Tagging results of arousal

	EEG only		Video only		Independent feature-fusion level	
	High	Low	High	Low	High	Low
High	125	24	46	24	114	35
Low	31	17	13	9	18	30
Accuracy	72.08 %		59.78 %		73.10 %	
F1-score	0.3820		0.3273		0.5310	
Average precision	0.5965		0.5331		0.6591	
	Decision-level fusion		Dependent feature-level fusion			
	High	Low	High	Low		
High	123	26	136	13		
Low	27	21	34	14		
Accuracy	73.10 %		76.14 %			
F1-score	0.4421		0.3733			
Average precision	0.6315		0.6022			

5.3.2 Tagging results of arousal

Table 2 shows the tagging results of arousal. The performance of tagging using EEG only is better than that using video only, since the former achieves higher scores in all the three parameters. However, both methods tend to misclassify samples of low arousal to high arousal. Independent feature-level fusion and decision -level fusion demonstrate their effectiveness by a higher F1-score when compared with tagging using one modality. Compared to tagging using video content, dependent feature-level fusion performs better, with higher accuracy, F1-score, and average precision. It further confirms the physiological signals' potential to bridge the semantic gap.

From the tagging results of valence and arousal, we can find that the independent feature-level fusion achieves the best performance among the three fusion methods, followed by decision-level fusion and dependent feature-level fusion. We used 197 samples in our fusion experiments. One video may have different tags if we infer the tag from video only. Since we first infer video tags from EEG and videos respectively during decision-level fusion, the issue of multiple tags for one video may impact the tagging performance from videos, thus reducing the performance of decision-level fusion. In independent feature-level fusion, we do not face such issue. This may be the reason that independent feature-level fusion performs better than decision-level fusion. A possible explanation for better performance of independent feature-level fusion as compared to dependent feature-level fusion may be that the distribution of EEG data may not satisfy linear Gaussian, which is the assumption of the dependent feature-level fusion model.

6 Conclusion

In this paper, we first propose a general framework of emotional tagging, consisting of video content space, physiological space, and emotional tag space and their relationships. Secondly, we design and conduct the psychophysiological experiment to collect data. Thirdly, three spaces are constructed and their relations are analyzed based on our collected data. Lastly, three emotion tagging approaches are proposed to annotate videos with emotional tags by combining video content and physiological signals, i.e., independent feature-level fusion, decision-level fusion, and dependent feature-level fusion.

The experimental results show that in valence and arousal spaces, hybrid methods produce better tagging performance than that of the conventional emotional tagging methods that use either video or EEG features alone. The independent feature-level fusion method achieves the best performance among the three hybrid methods.

All the recruited subjects in our psychophysiological experimental are students from our university, where the ratio of male/female is about 7:1. Thus, the age range of the subjects is narrow, and the ratio of male to female is unbalanced. We will enhance subject diversity during future data collection.

Recent research has demonstrated the gender difference in response to emotional stimuli [84]. Such gender differences may affect emotional video tagging. We will study this issue in the next step of our work.

With the growing availability of built-in sensors, we believe emotional tagging with the help of spontaneous user responses will attract more and more attention. However, not everybody is comfortable with wearing sensors to detect their body changes, or being observed by cameras during the actual tagging. A potential solution is to employ users' spontaneous responses during model training only. In actual tagging, only video features would be used, without requiring user response. This is a study of how to learn using privileged information [74]. We will investigate this study in the future.

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 ll(1106c0805008) and Youth Creative Project of USTC.

Thanks to all the volunteers for the participation in our experiments.

Appendix

A Video information

Table 3 Source video, type, start-end time and resolution

Source video names	Video type	Start-end time (HH:MM:SS)	Resolution
Shutter (Directed by Parkpoom Wongpoom and Banjong Pisonthanakun)	Movie	00:29:07–00:31:40	640 × 352
Shutter (Directed by Parkpoom Wongpoom and Banjong Pisonthanakun)	Movie	00:38:01–00:42:42	640 × 352
Dragons World A Fantasy Made Real	Animated film	01:07:49–01:10:10	672 × 352
Titanic	Movie	01:49:21–02:52:09	800 × 336
Leon - The Professional	Movie	00:03:39–00:07:54	1024 × 576
I Am Legend	Movie	00:26:23–00:31:30	640 × 360
Backkom Season 1:Episode 38 (Railway obstacles)	Animated film	00:00:04–00:03:01	1024 × 576

Table 3 (continued)

Source video names	Video type	Start-end time (HH:MM:SS)	Resolution
The Silence of the Lambs	Movie	01:14:33–01:18:25	720 × 456
The Day After Tomorrow	Movie	00:48:11–00:51:09	752 × 416
From Hell	Movie	01:42:32–01:45:03	1024 × 576
The Shining	Movie	01:48:02–01:53:46	640 × 480
Garfield	Animated film	00:23:41–00:28:06	672 × 368
Chaos Theory	Movie	00:33:28–00:38:33	640 × 360
The Extra-Terrestrial	Movie	01:33:53–01:38:07	720 × 480
The Grudge 1	Movie	00:27:19–00:31:58	512 × 384
The Grudge 1	Movie	00:49:00–00:54:13	512 × 384
The Grudge 2	Movie	00:44:53–00:50:01	592 × 320
The Grudge 2	Movie	01:21:12–01:23:07	592 × 320
The Uninvited	Movie	00:27:03–00:29:21	624 × 352
The Lion King	Animated film	00:13:53–00:18:11	544 × 320
The Lion King	Animated film	01:23:02–01:23:46	544 × 320
Monsters Inc	Animated film	00:57:33–01:02:40	800 × 432
Bee Movie	Animated film	00:13:33–00:16:13	720 × 396
Horton Hears a Who	Animated film	00:32:05–00:36:09	1280 × 720
Horton Hears a Who	Animated film	00:02:19–00:05:07	1280 × 720
Mountain Patrol	Movie	01:14:56–01:20:31	640 × 272
Braveheart	Movie	02:42:32–02:46:51	640 × 272
Braveheart	Movie	00:53:30–00:55:32	640 × 272
The Lion King	Animated film	01:01:23–01:04:06	544 × 320
Tarzan	Animated film	00:22:42–00:25:17	672 × 384
Cars	Animated film	00:00:41–00:03:19	1024 × 576
Bunhongshin	Movie	00:48:14–00:50:48	640 × 272
Discovery	Movie	00:58:32–01:02:10	640 × 360
Happy Feet	Animated film	00:25:04–00:27:24	640 × 360
Bunhongshin	Movie	01:11:02–01:13:01	640 × 272
The Ring (Directed by Hideo Nakata)	Movie	00:27:04–00:29:25	608 × 344
Face	Movie	00:24:30–00:26:32	576 × 304
The Grudge 1	Movie	00:13:31–00:15:32	640 × 360
The Hitchhiker's Guide to the Galaxy	Movie	00:58:30–01:00:12	1280 × 720
Twilight	Movie	00:49:44–00:53:43	1024 × 576
Twilight	Movie	01:37:23–01:40:29	1024 × 576
Braveheart	Movie	00:41:26–00:45:29	1024 × 560
Mother Love Me Once Again	Movie	00:28:27–00:30:57	640 × 432
Mother Love Me Once Again	Movie	01:03:21–01:08:41	640 × 432
Mother Love Me Once Again	Movie	01:08:56–01:13:07	640 × 432
Tears Of The Sun	Movie	01:29:13–01:34:38	800 × 432
Bolt	Animated film	01:22:20–01:24:52	640 × 336
Exorcist: The Beginning	Movie	00:36:42–00:39:54	640 × 272
High School Musical	Movie	01:46:00–01:50:07	880 × 480
Rob-B-Hood	Movie	01:08:05–01:13:13	640 × 272
Meet the Fockers 2	Movie	00:13:26–00:17:21	576 × 324
Red Cliff	Movie	00:13:45–00:17:16	672 × 272
Cheaper by the Dozen 2	Movie	00:09:07–00:12:32	800 × 452
The Sixth Sense	Movie	01:15:50–01:17:24	1280 × 720
The Thing	Movie	01:14:57–01:17:46	1024 × 576

Table 3 (continued)

Source video names	Video type	Start-end time (HH:MM:SS)	Resolution
Starship Troopers	Movie	01:48:34–01:50:09	1024 × 552
Starship Troopers	Movie	01:48:34–01:49:29	1024*552
Starship Troopers	Movie	01:49:30:01:50:05	1024*552
Slither	Movie	00:05:31–00:06:23	448*336
Slither	Movie	01:15:23–0:16:42	576*320
Slither	Movie	01:17:49–01:18:23	576*320
Slither	Movie	01:19:41–01:21:23	576*320
The Grudge 1	Movie	00:31:07–00:31:46	512*384
The Grudge 1	Movie	00:53:05–00:54:18	512*384
Shutter (Directed by Masayuki Ochia)	Movie	00:13:06–00:14:26	640*352
Shutter (Directed by Masayuki Ochia)	Movie	00:70:03–00:70:53	640*352
The Grudge 2	Movie	00:47:53–00:49:52	592*320
The Grudge 2	Movie	01:22:18–01:24:07	592*320
The Grudge 3	Movie	01:18:57–01:20:45	640*352
The Grudge 3	Movie	00:35:03–00:36:51	640*352
Bunhongsin	Movie	00:04:09–00:05:57	640*272
Bunhongsin	Movie	00:48:44–00:50:41	640*272
The Ring (Directed by Gore Verbinski)	Movie	00:27:04–00:29:02	608*344
The Ring (Directed by Gore Verbinski)	Movie	01:26:00–01:27:53	608*344
My Beloved	Movie	00:28:27–00:30:57	640*432
My Beloved	Movie	01:03:21–01:05:36	640*432
My Beloved	Movie	01:05:36–01:08:41	640*432
My Beloved	Movie	01:08:56–01:11:59	640*432
Garfield 2	Movie	01:05:12–01:06:06	800*432
Garfield	Movie	00:30:07–00:31:18	672*368
Garfield	Movie	00:26:08–00:27:41	672*368
iPartment Season 1:Episode 11	TV play	00:03:17–00:04:27	640*480
iPartment Season 1:Episode 11	TV play	00:04:29–00:06:21	640*480
iPartment Season 1:Episode 13	TV play	00:32:57–00:34:54	640*480
Mr. Bean Episode 4: Mr. Bean goes to town	TV play	00:20:09–00:21:44	512*384
Tom and Jerry Episode 1 West cowboy	TV play	00:01:36–00:03:39	464*316
Tom and Jerry Episode 6 The Three Little Kittens	TV play	00:00:25–00:01:45	464*316
Tom and Jerry Episode 6 The Three Little Kittens	TV play	00:02:21–00:04:03	464*316
Tom and Jerry Episode 9 bodyguard	TV play	00:01:06–00:03:06	464*316
Tom and Jerry Episode 90 the mouse from the hunger	TV play	00:00:22–00:02:07	400*300
Tom and Jerry Episode 90 the mouse from the hunger	TV play	00:02:35–00:03:38	400*300
Tom and Jerry Episode 90 the mouse from the hunger	TV play	00:04:02–00:06:12	400*300

References

1. AlZoubi O, Calvo RA, Stevens RH (2009) Classification of eeg for affect recognition: an adaptive approach. In: Nicholson A, Li X (eds) AI 2009: advances in artificial intelligence, vol 5866. Lecture notes in computer science. Springer Berlin Heidelberg, pp 52–61
2. 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. ACM, pp 461–470
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: IEEE International conference on multimedia and expo, ICME 2009. IEEE, pp 1440–1443
4. Arifin S, Cheung PYK (2006) User attention based arousal content modeling. In: IEEE international conference on image processing, pp 433–436
5. 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 semantic computing, ICSC 2007. IEEE, pp 147–154
6. Arifin S, Cheung PYK (2007) A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. In: Proceedings of the 15th international conference on Multimedia, MULTIMEDIA '07. ACM, New York, NY, USA, pp 68–77
7. Arroyo I, Cooper DG, Burleson W, Woolf BP, Muldner K, Christopherson R (2009) Emotion sensors go to school. In: Proceeding of the 2009 conference on artificial intelligence in education, July 6th–10th. Brighton, UK, IOS Press, pp 17–24
8. Bailenson JN, Pontikakis ED, Mauss IB, Gross JJ, Jabon ME, Hutcherson CAC, Nass C, John O (2008) Real-time classification of evoked emotions using facial feature tracking and physiological responses. Int J Human-Comput Stud 66(5):303–317
9. Banich MT, Compton RJ (2010) Cognitive neuroscience. Wadsworth Publishing Company
10. Bänziger T, Didier G, Scherer KR (2009) Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert). Emotion 9(5):691
11. Busso C, Deng Z, Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th international conference on multimodal interfaces, ICMI '04. ACM, New York, NY, USA, pp 205–211
12. Calvo RA, Brown I, Scheding S (2009) Effect of experimental factors on the recognition of affective mental states through physiological measures. In: Nicholson A, Li X (eds) AI 2009: advances in artificial intelligence, vol 5866. Lecture notes in computer science. Springer Berlin Heidelberg, pp 62–70
13. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1(1):18–37
14. Canini L, Gilroy S, Cavazza M, Leonardi R, Benini S (2010) Users' response to affective film content: a narrative perspective. In: International workshop on content-based multimedia indexing, (CBMI) 2010. IEEE, pp 1–6
15. Canini L, Benini S, Leonardi R (2013) Affective recommendation of movies based on selected connotative features. IEEE Trans Circuits Syst Video Technol 23(4):636–647
16. Cardani D (2001) Adventures in hsv space. Laboratorio de Robótica, Instituto Tecnológico Autónomo de México
17. Caridakis G, Karpouzis K, Kollias S (2008) User and context adaptive neural networks for emotion recognition. Neurocomputing 71(13–15):2553–2562
18. Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction, vol 4868. Lecture notes in computer science. Springer Berlin Heidelberg, pp 92–103
19. Chan CH, Jones GJF (2005) Affect-based indexing and retrieval of films. In: Proceedings of the 13th annual ACM international conference on multimedia, MULTIMEDIA '05. ACM, New York, NY, USA, pp 427–430
20. Chuang Z-J, Wu C-H (2004) Multi-modal emotion recognition from speech and text. Comput Linguist Chin Lang Process 9(2):45–62
21. D'Mello SK, Graesser A (2010) Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model User-Adapt Interact 20:147–187

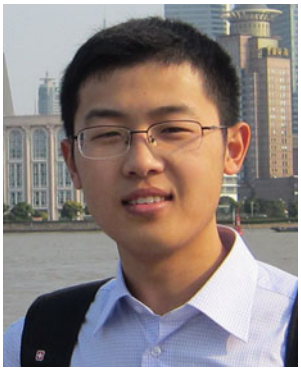
22. Haag A, Goronzy S, Schaich P, Williams J (2004) Emotion recognition using bio-sensors: first steps towards an automatic system. In: Affective dialogue systems, vol 3068. Lecture notes in computer science. Springer Berlin Heidelberg, pp 36–48
23. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimedia 7(1):143–154
24. Hanjalic A (2006) Extracting moods from pictures and sounds: towards truly personalized tv. IEEE Signal Process Mag 23(2):90–100
25. Heraz A, Frasson C (2007) Predicting the three major dimensions of the learner's emotions from brainwaves. World Acad Sci Eng Technol 25:323–329
26. Hussain S Md, Calvo RA, Pour PA (2011) Hybrid fusion approach for detecting affects from multichannel physiology. In: D'Mello S, Graesser A, Schuller B, Martin J-C (eds) Affective computing and intelligent interaction, vol 6974. Lecture notes in computer science. Springer Berlin Heidelberg, pp 568–577
27. Irie G, Hidaka K, Satou T, Kojima A, Yamasaki T, Aizawa K (2009) Latent topic driving model for movie affective scene classification. In: Proceedings of the 17th ACM international conference on multimedia, MM '09. ACM, New York, NY, USA, pp 565–568
28. Izard CE, Kagan J (1988) Emotions, cognition, and behavior. Cambridge Univ Pr
29. Ji Z, Qin S (2003) Detection of eeg basic rhythm feature by using band relative intensity ratio (brir). In: Proceedings IEEE international conference on acoustics, speech, and signal processing, (ICASSP'03), vol 6. IEEE, pp VI–429
30. Joho H, Staiano J, Sebe N, Jose JM (2011) Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. Multimed Tools Appl 51:505–523
31. Kaliouby R, Robinson P (2005) Generalization of a vision-based computational model of mind-reading. In: Tao J, Tan T, Picard RW (eds) Affective computing and intelligent interaction, vol 3784. Lecture notes in computer science. Springer Berlin Heidelberg, pp 582–589
32. Kang HB (2003) Affective contents retrieval from video with relevance feedback. Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. Lecture Notes in Computer Science 2911:243–252
33. Kang HB (2003) Affective content detection using hmms. In: Proceedings of the eleventh ACM international conference on multimedia, MULTIMEDIA '03. ACM, New York, NY, USA, pp. 259–262
34. Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environments. In: Proceedings of the 13th annual ACM international conference on multimedia, MULTIMEDIA '05. ACM, New York, NY, USA, pp 677–682
35. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Human-Comput Stud 65(8):724–736
36. Karpouzis K, Caridakis G, Kessous L, Amir N, Raouzaiou A, Malatesta L, Kollias S (2007) Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In: Huang TS, Nijholt A, Pantic M, Pentland A (eds) Artifical intelligence for human computing, vol 4451. Lecture notes in computer science. Springer Berlin Heidelberg, pp 91–112
37. Kemp AH, Gray MA, Eide P, Silberstein RB, Nathan PJ (2002) Steady-state visually evoked potential topography during processing of emotional valence in healthy subjects. NeuroImage 17(4):1684–1692
38. Kensinger EA (2004) Remembering emotional experiences: the contribution of valence and arousal. Reviews in the Neurosci 15:241–252
39. Kim J, Andr E (2006) Emotion recognition using physiological and speech signal in short-term observation. In: Perception and interactive technologies, vol 4021. Lecture notes in computer science. Springer Berlin Heidelberg, pp 53–64
40. Kim J (2007) Bimodal emotion recognition using speech and physiological changes. In: Grimm M, Kroschel K (ed) Robust speech recognition and understanding. I-Tech Education and Publishing, Vienna, Austria, pp 265–280. ISBN 978-3-902613-08-0
41. Knautz K, Wolfgang GS (2011) Collective indexing of emotions in videos. J Doc 67(6):975–994
42. Koelstra S, Mühl C, Patras I (2009) Eeg analysis for implicit tagging of video data. In: 3rd International conference on affective computing and intelligent interaction and workshops, ACII 2009. IEEE, pp 1–6
43. 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: Brain informatics, vol 6334. Lecture notes in computer science. Springer Berlin Heidelberg, pp 89–100

44. Koelstra S, Mühl C, Soleymani M, Jong-Seok L, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2012) Deap: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput 3(1):18–31
45. Krolak-Salmon P, Hénaff. M-A, Vighetto A, Bertrand O, Mauguière F et al (2004) Early amygdala reaction to fear spreading in occipital, temporal, and frontal cortex: a depth electrode erp study in human. Neuron 42(4):665–676
46. Krzywicki AT, He G, O'Kane Bl (2009) Analysis of facial thermal variations in response to emotion: eliciting film clips. In: Proceedings of SPIE, vol 7343, p 734312
47. Kulic D, Croft E (2007) Affective state estimation for human-robot interaction. IEEE Trans Robot 23(5):991–1000
48. Kyung Hwan K, Seok Won Bg, Sang Ryong K (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42:419–427
49. Liu C, Conn K, Sarkar N, Stone W (2008) Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder. Int J Human-Comput Stud 66(9):662–677
50. Lu Y, Sebe N, Hytnen R, Tian Q (2011) Personalization in multimedia retrieval: a survey. Multimed Tools Appl 51(1):247–277
51. Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl 49:277–297
52. McLaren K (1976) Xiii - the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula. J Soc Dyers Colour 92(9):338–341
53. Molau S, Pitz M, Schlter R, Ney H (2001) Computing mel-frequency cepstral coefficients on the power spectrum. In: Proceedings IEEE International conference on acoustics, speech, and signal processing, (ICASSP '01), vol 1, pp 73–76
54. 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
55. Money AG, Agius H (2008) Feasibility of personalized affective video summaries. Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science 4868:194–208
56. Money AG, Agius H (2010) Elvis: entertainment-led video summaries. ACM Trans Multimedia Comput Commun Appl (TOMCCAP) 6(3):17
57. Muneesawang P, Guan L (2006) Multimedia database retrieval: a human-centered approach. Springer
58. Murphy KP (1998) Inference and learning in hybrid bayesian networks. University of California, Berkeley, Computer Science Division
59. Nasoz F, Alvarez K, Lisetti CL, Finkelstein N (2004) Emotion recognition from physiological signals using wireless sensors for presence technologies. Cogn Technol Work 6:4–14
60. 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, New York, NY, USA, pp 105–114
61. Ong KM, Wataru K (2009) Classification of video shots based on human affect. Inf Media Technol 4(4):903–912
62. Peng WT, Chu WT, Chang CH, Chou CN, Huang WJ, Chang WY, Hung YP (2011) Editing by viewing: automatic home video summarization by viewing behavior analysis. IEEE Trans Multimedia 13(3):539–550
63. Rasheed Z, Sheikh Y, Shah M (2005) On the use of computable features for film classification. IEEE Trans Circuits Syst Video Technol 15(1):52–64
64. Satoshi T, Takashi K (2009) Video abstraction method based on viewer's heart activity and its evaluations. J Inst Image Inf Telev Eng 63(1):86–94
65. Schaefer A, Nils F, Sanchez X, Philippot P (2010) Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cogn Emot 24(7):1153–1172
66. Sebe N, Cohen I, Gevers T, Huang TS (2005) Multimodal approaches for emotion recognition: a survey. Internet Imaging VI 5670:56–67
67. Smeaton AF, Rothwell S (2009) Biometric responses to music-rich segments in films: the cdvplex. In: Seventh international workshop on content-based multimedia indexing, 2009. CBMI'09. IEEE, pp 162–168
68. 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. ACM, pp 32–39

69. Soleymani M, Joep JMK, Guillaume C, Thierry P (2009) A bayesian framework for video affective representation. In: 3rd International conference on affective computing and intelligent interaction and workshops, ACII 2009, pp 1–7
70. Soleymani M (2011) Implicit and automated emotional tagging of videos, 11/04 2011. ID: unige:17629
71. Soleymani M, Koelstra S, Patras I, Pun T (2011) Continuous emotion detection in response to music videos. In: IEEE international conference on automatic face gesture recognition and workshops (FG 2011), pp 803–808
72. Sun K, Yu J (2007) Video affective content representation and recognition using video affective tree and hidden markov models. In: Paiva A, Prada R, Picard RW (eds) Affective Computing and Intelligent Interaction. Second International Conference, ACII 2007 Lisbon, Portugal, September 12–14, 2007 Proceedings. Lecture Notes in Computer Science, vol 4738. Springer, Heidelberg, pp 594–605
73. Teixeira RMA, Yamasaki T, Aizawa K (2012) Determination of emotional content of video clips by low-level audiovisual features. Multimed Tools Appl 61:21–49
74. Vapnik V, Vashist A (2009) A new learning paradigm: learning using privileged information. Neural Netw 22(5):544–557
75. Villon O, Lisetti C (2006) A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. In: The 15th IEEE international symposium on robot and human interactive communication, ROMAN 2006, pp 269–276
76. Vinciarelli A, Suditu N, Pantic M (2009) Implicit human-centered tagging social sciences. IEEE Signal Process Mag 26(6):173–180
77. Wagner J, Kim J, André E (2005) From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: IEEE International conference on multimedia and expo, ICME 2005, pp 940–943
78. Wang HL, Cheong LF (2006) Affective understanding in film. IEEE Trans Circuits Syst Video Technol 16(6):689–704
79. Wang CW, Cheng WH, Chen JC, Yang SS, Wu JL (2006) Film narrative exploration through the analysis of aesthetic elements. Adv Multimed Model. Lecture Notes in Computer Science 4351:606–615
80. Wang S, Wang X (2010) Emotional semantic detection from multimedia: a brief overview. In: Dai Y, Chakraborty B, Shi M (ed) Kansei engineering and soft computing: theory and practice. IGI Global, USA, pp 126–146. doi:10.4018/978-1-61692-797-4, ISBN13: 9781616927974, ISBN10: 1616927976, EISBN13: 9781616927998
81. Watanapa SC, Thipakorn B, Charoenkitarn N (2008) A sieving ann for emotion-based movie clip classification. IEICE Trans Inf Syst 91(5):1562–1572
82. 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, 2004. ICME'04. 2004, vol 2. IEEE, pp 831–834
83. Winoto P, Tang TY (2010) The role of user mood in movie recommendations. Expert Syst Appl 37(8):6086–6092
84. Wrase J, Klein S, Gruesser SM, Hermann D, Flor H, Mann K, Braus DF, Heinz A (2003) Gender differences in the processing of standardized emotional visual stimuli in humans: a functional magnetic resonance imaging study. Neurosci Lett 348(1):41–45
85. Xu M, Jin JS, Luo S, Duan L (2008) Hierarchical movie affective content analysis based on arousal and valence features. In: Proceedings of the 16th ACM international conference on multimedia. ACM, pp 677–680
86. Xu M, Wang J, He X, Jin JS, Luo S, Lu H (2012) A three-level framework for affective content analysis and its case studies. Multimed Tools Appl 1–23. doi:10.1007/s11042-012-1046-8
87. Xu M, Xu C, He X, Jin JS, Luo S, Rui Y (2012) Hierarchical affective content analysis in arousal and valence dimensions. Signal Process. In press
88. Yazdani A, Lee JS, 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. ACM, pp 81–88
89. Yoo HW, Cho SB (2007) Video scene retrieval with interactive genetic algorithm. Multimed Tools Appl 34(3):317–336
90. 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

91. 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
92. Zhao Y (2012) Human emotion recognition from body language of the head using soft computing techniques. PhD thesis, University of Ottawa

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.



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 Ph.D degree there. His research interest is Affective Computing.

Guobing Wu received the Bachelor degree in School of Computer Science and Technology from Anhui University, Hefei, Anhui Province, China, in 2009. And he received the M.S. degree in School of Computer Science and Technology from University of Science and Technology of China, Hefei, Anhui Province, China, in 2012. 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 Rensselaer Polytechnic Institute (RPI). He recently served as a program director at the National Science Foundation (NSF), where he managed NSF's computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, the Robotics Institute at Carnegie Mellon University, the Dept. of Computer Science at University of Nevada at Reno, and the US Air Force Research Laboratory. Prof. Ji currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI.

Prof. Ji's research interests are in computer vision, probabilistic graphical models, information fusion, and their applications in various fields. He has published over 160 papers in peer-reviewed journals and conferences. His research has been supported by major governmental agencies including NSF, NIH, DARPA, ONR, ARO, and AFOSR as well as by major companies including Honda and Boeing. Prof. Ji is an editor on several related IEEE and international journals and he has served as a general chair, program chair, technical area chair, and program committee member in numerous international conferences/workshops. Prof. Ji is a fellow of IAPR.