Neurocomputing 291 (2018) 11–20

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A novel feature set for video emotion recognition

Shasha Mo [a], Jianwei Niu [a,*], Yiming Su [a], Sajal K. Das [b]

[a] New Main Building, Beihang University, Beijing 100191, China
[b] Department of Computer Science and Engineering, University Texas at Arlington, Arlington, TX 76019, USA

ARTICLE INFO

Article history:
Received 23 January 2017
Revised 21 November 2017
Accepted 5 February 2018
Available online 16 February 2018

Communicated by Prof Fuxin Li

Keywords:
Video affective content analysis
Cross-correlation
Empirical mode decomposition
Hilbert–Huang transform
Feature extraction

ABSTRACT

In video recommendation systems, emotions are used along with several other proposed content-based video features. However, such features are independently based on visual or audio signals and the relationship representing the dependencies between the visual and the audio signals is still unexplored. In order to solve this problem, a novel feature set called HHTC features based on the combination of Hilbert–Huang Transform (HHT) based visual features, HHT-based audio features, and cross-correlation features is proposed in this paper. In addition to the dependencies between the visual and the audio signals, the proposed HHTC features have the ability to indicate the time-varying characteristics of these signals. The proposed features are applied to video emotion recognition with the Support Vector Regression (SVR) with potential use in video affective recommendation systems. Experimental results demonstrate that the proposed approach can achieve an improved performance of video affective recognition.

1. Introduction

Emotion is one of the main factors that influence daily life. Videos, as an information carrying tool, have played an increasing role over the past years [1,2]. Video emotion recognition is a method that combines the videos with human emotions and is defined as extracting the emotions of a viewer from his or her current affective state, social experiences and user profile. Video provides multimodal data in terms of vocal and visual modality, thus most work in this field focuses on multimodal emotion recognition by using visual and audio information [3]. It is believed that video emotion recognition can improve the performance of video recommendation systems. Compared with the traditional content-based video analysis, the emotional analysis is more user-centered because it takes more psychological basis into consideration [4–6].

Neurological studies have shown several emotion models representing different human emotions. Since the nature of the basic emotional units is essentially dimensional or categorical, the emotion models can be divided into two types [7,8]. In the *categorical emotion theories*, the emotions can be categorized into a limited number of distinct emotion types with each type having specific characteristic properties [9–11]. In contrast, the *dimensional emotion theories* conceptualize emotions as arising from combination of fundamental dimensions [12–14]. The emotion models are the basis of the emotion recognition procedure. The goal of emotion recognition is to fill the gap between human affective states and video structures. A proper machine learning algorithm is used to complete the mapping from human emotions to video features.

The affective features are extracted from the videos to effectively recognize the emotions from the videos. The extracted features will then be transformed to an appropriate format for further processing. The feature extraction has a significant effect on the result of video affective content analysis. Several feature extraction methods have been proposed in past years [15,16], but most of them have focused on independently extracting visual or audio features. The dependencies between visual and audio signals have not been captured. In film-making, the directors usually express the emotional cue by combining the shots with sounds that correspond to the visual and audio signals. The combination of these two factors could fully express the affective states conveyed by the videos.

In this paper, the Hilbert–Huang Transform (HHT) [17–19] combined with the cross-correlation technique [20,21] is proposed for feature extraction in video affective content analysis. The proposed features are called the HHTC features and contain three categories: the HHT-based visual features, the HHT-based audio features and the cross-correlation features. Each category of the proposed features has three components, including the instantaneous frequency, the instantaneous amplitude, and the instantaneous phase. The HHT-based visual and audio features can provide a time-frequency-energy description of videos. In addition, the cross-correlation features are proposed to represent the dependencies between video frames and sounds. The proposed features can

* Corresponding author.
E-mail address: niujianwei@buaa.edu.cn (J. Niu).

https://doi.org/10.1016/j.neucom.2018.02.052
0925-2312/

represent the initial information of videos and focus on improving the performance of the single-label video mood recognition. Experimental results demonstrate that the proposed features can achieve a relatively high performance in the video affective recognition system.

The main contributions of this paper can be summarized as follows:

(1) A shot segmentation algorithm is proposed to obtain the key frame of each shot before the feature extraction. Each video shot generally includes at least tens of frames, and the proposed features are extracted only from the obtained key frame, which effectively reduces the computational load.
(2) A novel feature set, called the HHTC features, is proposed for video affective recognition systems. In the proposed feature extraction, the HHT is used to individually extract HHT-based visual and audio features from the visual and the audio signals, respectively. The dependencies between these signals are represented using the cross-correlation technique that results in a series of cross-correlation features.
(3) Once the HHTC features are obtained, a feature selection algorithm is applied to each feature type to reduce the number of components. The selected features are then used for video affective content analysis.
(4) In order to bridge the gap between the low features and the affective states, the Support Vector Regression (SVR) [22,23] is applied to map the HHTC features to emotions of the viewers. Two databases, the Video Affective Content Analysis Database (VACAD) and the LIRIS-ACCEDE database [24,25], are employed for video affective content analysis.

The rest of this paper is organized as follows. Section 2 describes the related work, including the emotion model and the existing content-related features. The extraction of the HHTC features based on the HHT and the cross-correlation technique are presented in Section 3. The methodology of video affective recognition system is described in Section 4 followed by the empirical experiments and performance evaluations of the proposed algorithm in Section 5. Finally, conclusions are drawn in Section 6 with directions of future research.

2. Related work

Up till now, considerable attention has been paid to video affective content analysis and applications, and many related works have been reported. In this section, the emotion models are introduced first, and then several existing emotional-related features are presented.

2.1. Emotion models

The emotion models used in videos and emotion research can be divided into two categories as the dimensional approach and the categorical approach. In the dimensional approach, the human emotions are conceptualized by defining their positions in two or sometimes three dimensions. Several dimensional models of emotion have been proposed in the literature and most of them incorporate valence and arousal or intensity dimensions [26–30]. A widely used two-dimensional model is the Russell's emotion model [14] that consists of two dimensions: valence and arousal as illustrated in Fig. 1. All emotions in this model have varying degrees of valence and arousal, and are distributed in a two-dimensional circular space. Arousal is the vertical axis and varies form "uprising" or "exacted" to "silent" or "calm". Valence is the horizontal axis and indicates the affective state from "positive" or "pleasant" to "negative" or "unpleasant". In another point of view, the arousal is the intensity of emotion, while the valence

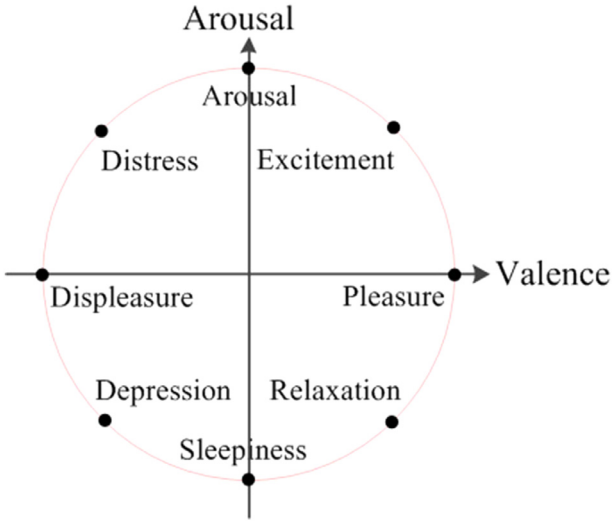

Fig. 1. Russell's emotion model [14].

Table 1
Emotion categories proposed by Hu and Downie [10].

Cluster	Mood
I	Rousing, Passionate, Confident, Boisterous, Rowdy
II	Cheerful, Rollicking, Fun, Sweet, Amiable/good natured
III	Poignant, Literate, Wistful, Bittersweet, Autumnal, Brooding
IV	Hyperbolic tangent & Silly, Humorous, Campy, Quirky, Whimsical, Witty, Wry
V	Fiery, Aggressive, Tense/Anxious, Intense, Volatile, Visceral

is the type of emotion. The center of the circle in Russell's emotion model represents a medium level of arousal and valence. The other available dimensional models include the vector model [26], the Positive Activation–Negative Activation (PANA) model [27], the Plutchik's model [28], the PAD emotional state model [29] and the Lovheim cube of emotion [30].

In addition to the dimensional approach, all humans are seen to have an inherent set of basic emotions that are different from each other within the categorical view. These basic emotions are distinguishable by the facial expression of an individual and biological process. Six basic emotions, including anger, fear, happiness, sadness, disgust, and surprise have been concluded by a cross-cultural study of Ekman in 1972 [9]. Each emotion in the Ekman's model acts not as a discrete category instead of an individual emotional state. Hevner's affective checklist [31] is another categorical approach that describes eight clusters of affective states as shown in Fig. 2. Each cluster has several similar elements and lays out in a cumulative way according to their meaning. Hu and Downie [10] also proposed a model consisting of five emotion categories as presented in Table 1. Among all the categorical models, Ekman's model is used for video affective content analysis in this paper.

2.2. Emotion-related feature extraction

A video consists of two main types of data: visual data and auditory data. Generally, both the visual and the auditory contents of videos are carefully designed by the artists to coordinate with each other. The extraction of suitable features that effectively characterize the different affective contents plays an essential role in the video affective content analysis. Some studies have shown that certain visual or audio features are related to the emotion of a video. In this subsection, the visual and the audio features used for video affective content analysis are reviewed.

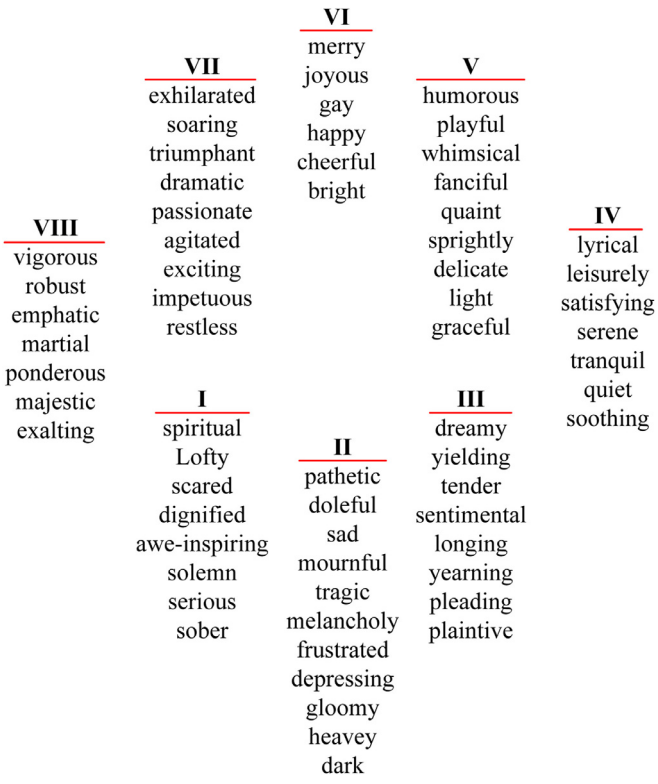

Fig. 2. Hevner's affective checklist [31].

2.2.1. Audio features

The audio features are informative in affective content characterization. Generally, the audio signal of a video is a mixture of sounds from different sources, including speech, music, and environmental sound. Numerous research studies have been conducted on audio feature extraction due to the importance of audio features in characterizing the emotion of videos. Typical audio features include zero crossing rate, sound energy, pitch, onset, beat strength, tempo, bandwidth, spectral centroid, spectral rolloff, spectral flux, loudness, and Mel-frequency Cepstrum Coefficients (MFCC) et al. [12,32,33]. Different audio features can capture different types of emotions. For example, zero crossing rate, sound energy, pitch, onset, beat strength, loudness, and tempo are effective in classifying high or low arousal, while some other audio features are commonly related to valence.

2.2.2. Visual features

The visual features are extracted from the visual signal of a video, and can be used to reverse the intent of directors and the affective content of a video. Various techniques have been proposed to obtain the visual features of videos. Based on the existing feature extraction technology, visual features mainly include lighting, saturation, color energy, color heat, shot-change rate, shot duration, shot type transition rate, motion intensity, and motion dynamics et al. [11,12,34]. Tempo has a significant role in attracting viewers attention and in affecting reviewers emotion. Shot and motion are the two key elements that can control the tempo of a video, and the corresponding shot-related features and motion-related features have the ability of capturing the emotions of a video. Lighting and color are also connotative signatures that can affect the reviewers emotion and are used in movies scenes to create affective effects. The lighting-related and the color-related features have been used for characterizing the affective content of videos.

3. HHTC feature extraction

The feature extraction is an important step in the video affective content analysis. The aim is to extract the initial information that will influence the affective state of viewers. This section presents the HHTC feature extraction, which is based on the HHT and the cross-correlation technique. The proposed features have the ability of covering the visual and the audio signals along with their dependencies. Both sounds and frames in videos can be seen as a data set to which the HHT algorithm can be applied. The cross-correlation technique is a measure of similarity of two variables or vectors, which is a function of the lag of one relative to the other. In this paper, the correlation analysis is used to explore the relationship between frames and sounds in videos that will produce a series of features for video affective content analysis. In order to obtain the HHTC features, the HHT model is introduced first, followed by the processing of the HHTC features using the HHT and the cross-correlation technique. The proposed HHTC features include three types: the HHT-based visual features, the HHT-based audio features and the cross-correlation features.

3.1. HHT model

The HHT technique is designed specifically as an alternative method to provide a time-frequency-energy description of time varying data. It is a combination of a decomposition method called Empirical Mode Decomposition (EMD), and a spectral analysis tool called Hilbert Spectral Analysis (HSA). The EMD method proposed by Huang et al. [35]. decomposes any complicated data set into individual characteristic oscillatory modes known as Intrinsic Mode Functions (IMFs). The HSA method is applied to the IMFs to obtain the Hilbert spectrum of the input data.

An IMF is defined with the help of following two requirements:

(1) In the entire time sequence of input data, the number of extrema and zeros-crossing must be equal or differ at most by one.
(2) The mean value of the envelope defined by the local maxima and the local minima, is zero at any point.

With the above two requirements of IMFs, any given signal $x(t)$ can be decomposed by a sifting process. In order to extract an IMF from $x(t)$, all the local extrema, including maxima and minima, should firstly be identified. Then cubic spline lines are used as the upper and the lower envelopes by connecting all the local maxima and the minima, respectively. The upper envelope combined with the lower envelope should cover all the data. Finally, the first component is obtained by the following formula:

$$d(t) = x(t) - \frac{x_{up}(t) + x_{low}(t)}{2} \tag{1}$$

where $x_{up}(t)$ and $x_{low}(t)$ are the envelope and lower envelopes, respectively. Ideally, the component computed by Eq. (1) should satisfy the above-listed two requirements of IMFs even though the computed component may not have the features suitable for IMFs. In order to solve this problem, the operation as Eq. (1) will be repeated after the first round of sifting. In the subsequent sifting process, the component of kth iteration can be calculated by

$$d_{1m}(t) = d_{1(m-1)}(t) - \frac{d_{1(m-1),up}(t) + d_{1(m-1),low}(t)}{2}, \quad for\ m > 1 \tag{2}$$

where $d_{1(m-1),up}(t)$ and $d_{1(m-1),low}(t)$ are the upper and lower envelopes of the $m-1$ component. As this sifting process continues, the IMF requirements are checked with each iteration until the component $d_{1m}(t)$ is a monotonous function. The final component is then designed as the first IMF component of the data:

$c_1(t) = d_{1m}(t)$. In general, the first IMF component $c_1(t)$ shown in Eq. (3) should contain the shortest period component of the signal $x(t)$. The residue is $r_1(t) = x(t) - c_1(t)$ containing longer period variations in $x(t)$ is obtained by separating $c_1(t)$ from the rest of the signal. The residue $r_1(t)$ will then be treated as a new signal in the same sifting process. Thus, the IMFs embedded in the signal are extracted one by one with the sifting process. Finally, the original signal $x(t)$ can be reconstructed as follows

$$x(t) = \sum_{k=1}^{N} c_k(t) + r_N(t) \tag{3}$$

where N is the number of IMFs. Eq. (3) shows a N level decomposition of the original signal. The decomposition process is a scale filtering process, and each IMF reflects the characteristic scale of the signal and represents the inherent modal characteristics.

After obtaining the IMFs of the original signal, the second step is the HSA using the Hilbert transform defined by Cauchy Principal Value (CPV) [36] that leads to an apparent time-frequency description of signals. In the second step of the HHT, the Hilbert transform is applied to each IMF component and is given by:

$$H[c_k(t)] = PV \cdot \frac{1}{\pi t} \otimes c_k(t) = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{c_k(\tau)}{t-\tau} d\tau \tag{4}$$

where $\otimes$ denotes the convolution operation, and PV indicates the CPV. Then the analytic function $Z(t)$ is obtained by taking $c_k(t)$ as the real part and $H[c_k(t)]$ as the imaginary part:

$$Z_k(t) = c_k(t) + jH[c_k(t)] = a_k(t) exp[j\theta_k(t)] \tag{5}$$

where

$$a_k(t) = \sqrt{c_k^2(t) + H[c_k(t)]^2} \tag{6}$$

and

$$\theta_k(t) = \arctan \frac{H[c_k(t)]}{c_k(t)} \tag{7}$$

are the instantaneous amplitude and the instantaneous phase, respectively. The instantaneous frequency can then be extracted with the derivative of the phase:

$$\omega_k(t) = \frac{d\theta_k(t)}{dt} \tag{8}$$

After the IMF components and the instantaneous frequency are obtained, the original signal can be expressed as the following form:

$$x(t) = Re\left\{ \sum_{k=1}^{N} a_k(t) exp\left[k \int \omega_k(t) dt \right] \right\} \tag{9}$$

where $Re\{\cdot\}$ is the real part of $\{\cdot\}$. Here Eq. (7) can be seen as a generalized form of the Fourier transform.

The original HHT is used to analyze one-dimensional signal. In order to process two-dimensional signals, a two-dimensional HHT is constructed, in which both the EMD and the Hilbert transform are extended to two-dimensional case. Compared with the original EMD, the two-dimensional EMD identifies the extrema of the input two-dimensional signal by sliding a 3-by-3 gird and generates the upper and lower surfaces according to the maxima and the minima, respectively. The local means of the upper and lower surfaces are then computed, and the sifting process is applied. The input two-dimensional signal can be reconstructed with the two-dimensional EMD in the following form:

$$s(x, y) = \sum_{k=1}^{N} c_k(x, y) + r_N(x, y) \tag{10}$$

where $c_k(x, y)$ is the kth IMF component, and $r_N(x, y)$ is the residual. Finally, the obtained two-dimensional IMFs are transformed into the time-frequency domain using a two-dimensional Hilbert transform:

$$H[c_k(x, y)] = PV \cdot \frac{1}{\pi^2 xy} \otimes c_k(x, y) \tag{11}$$

Similar to the one-dimensional case, the corresponding instantaneous amplitude, the instantaneous phase and the instantaneous frequency can be obtained by Eq. (11):

$$a_k(x, y) = \sqrt{c_k^2(x, y) + H[c_k(x, y)]^2}, \tag{12}$$

$$\theta_k(x, y) = arc \tan \frac{H[c_k(x, y)]}{c_k(x, y)}, \tag{13}$$

$$\omega_k(x, y) = \frac{d\theta_k(x, y)}{dxdy}. \tag{14}$$

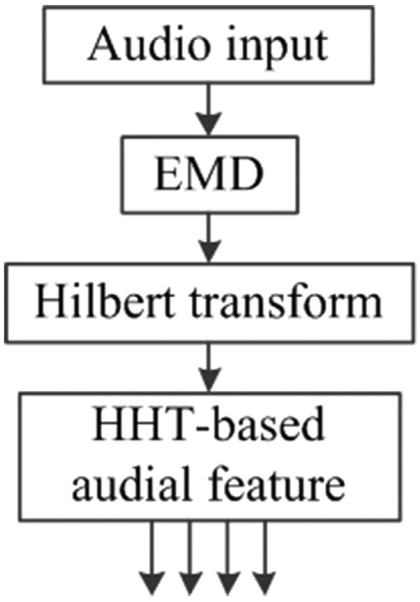

Fig. 3. The HHT-based audio feature extraction algorithm.

3.2. HHT-based audio features

The block diagram of the HHT-based audio feature extraction is shown in Fig. 3. The audio signal from a video is first iteratively decomposed into N IMF components along with the audio features shown in Eq. (3). Each IMF component is then transformed into the time-frequency domain using the Hilbert transform. Finally, the HHT-based audio features, such as the instantaneous frequency, the instantaneous amplitude and the instantaneous phase, are extracted from the Hilbert transform of the IMFs. The IMF components do not convey any information on the local properties, and the Hilbert transform is used for the calculation of HHT-based audio features.

In the HHT-based audio feature extraction, the number of IMFs will directly affect the feature vector size influencing the overall computation time. The selection criteria of the IMFs for a particular audio signal is the same as the termination condition of the sifting process. The Standard Deviation (SD) between two consecutive iterations is taken as the termination condition and computed as follows:

$$SD = \sum_{t=0}^{T} \frac{|d_{k-1}(t) - d_k(t)|^2}{d_{k-1}^2(t)} \tag{15}$$

where k is the number of the sifting process, and T is the length of the input signal. Generaly, different signal lengths have different SD values.

3.3. HHT-based visual features

Each video frame is a TrueColor image and stored as an m-by-n-by-3 data set that defines red, green and blue color components for each individual pixel. In order to extract the HHT-based visual features, the two-dimensional EMD method is applied to each color plane to obtain the IMFs [37]. After the two-dimensional EMD method, the original image can be decomposed into a sum of IMFs. The instantaneous frequency, the instantaneous amplitude and the instantaneous phase of each IMF are extracted with the two-dimensional Hilbert transform.

The two-dimensional EMD method is completely a data-driven approach and does not depend on the choice of the filter. This method is especially suitable for non-linear signals, for example texture images, and can better reflect the internal information of images. Similar to the one-dimensional case, the number of IMF component obtained from the two-dimensional EMD has an effect on the visual features. A termination condition for the two-dimensional case is given by [38]:

$$SD = \sum_{x=0}^{N_1}\sum_{y=0}^{N_2}\frac{|d_{k-1}(x,y)-d_k(x,y)|^2}{d_{k-1}^2(x,y)} \tag{16}$$

where SD is the standard deviation between two consecutive sifting processes of the two-dimensional case; $d_k(x, y)$ is the component of the two-dimensional sifting process; N_1 and N_2 are respectively the size of the input video frames and k is the number of sifting.

3.4. Cross-correlation features

This subsection introduces the cross-correlation features that represent the dependencies between the audio and the visual signals. The cross-correlation features proposed in this paper are computed with the audio signal and the motion vector, and motion vectors can track the moving objects in video image sequences. The motion vectors are first computed from video frame sequences and then a cross-correlation between the audio signal and the motion vectors is applied to obtain a correlation function. The cross-correlation function between two signals is calculated as follows [20]:

$$R_{av}(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_0^T a(t)v(t+\tau)dt \tag{17}$$

where T is the signal length, τ is the argument of cross-correlation function, and $a(t)$ and $v(t)$ are the audio signal and the motion vector, respectively. The computed cross-correlation function is transformed into an envelope by using the Hilbert transform as:

$$H[R_{av}(\tau)] = \widetilde{R}_{av}(\tau) = \frac{1}{\pi}\int_{-\infty}^{\infty}R_{av}(\tau')\frac{1}{\tau-\tau'}d\tau'. \tag{18}$$

After the Hilbert transform, a correlation spectra is obtained and applied to the cross-correlation feature extraction. The cross-correlation feature has three components: the instantaneous amplitude, the instantaneous phase, and the instantaneous frequency. The block diagram of the cross-correlation feature extraction is shown in Fig. 5.

4. Methodology

This section provides an overview of the affective recognition system. In the training stage, the proposed HHTC features are first extracted using the HHT and the cross-correlation technique from videos. Then a feature selection algorithm is applied to select the most relevant HHTC features to the videos. Finally, a machine learning algorithm is used to map the extracted features to the six emotions proposed by Ekman, which includes anger, fear, happiness, sadness, disgust, and surprise. A model for bridging the gap between human affective states and video structures is established. In the testing stage, the same features as in the training stage are extracted from the video clips, and the same procedure is applied to obtain the emotion of testing clips. The framework of the proposed affective video recognition system is shown in Fig. 6. All the steps of the proposed framework are described in the following subsections.

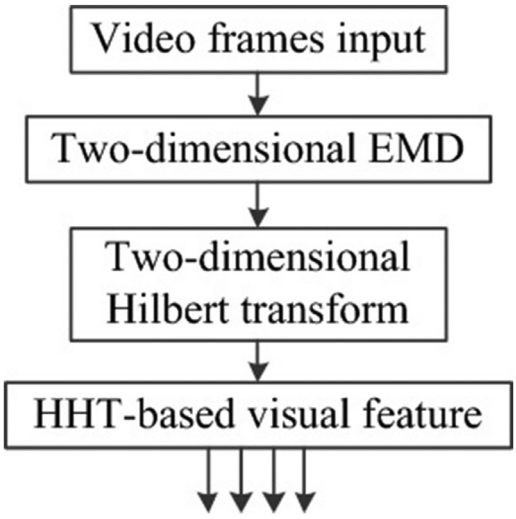

Fig. 4. The HHT-based visual feature extraction algorithm.

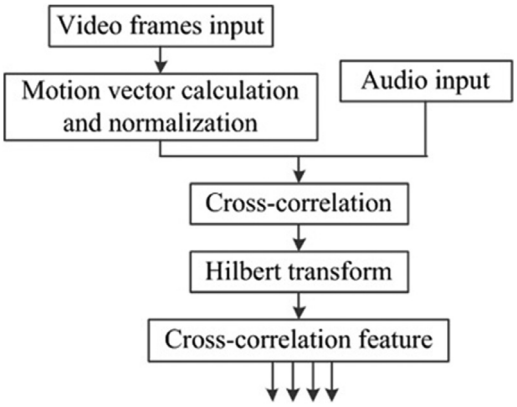

Fig. 5. The cross-correlation feature extraction algorithm.

4.1. Related techniques for feature extraction

Affective feature extraction is an important step for video affective recognition systems, and this subsection introduces the proposed HHTC features. The visual and the audio contents are carefully designed by artists to express the emotions to the viewers, thus both are used for the proposed feature extraction. In the feature extraction step, the visual and the audio signals are first proposed separately for extracting the visual and the audio features using the HHT, respectively. Then both the visual and the audio signals are used to compute the cross-correlation feature using the cross-correlation technique.

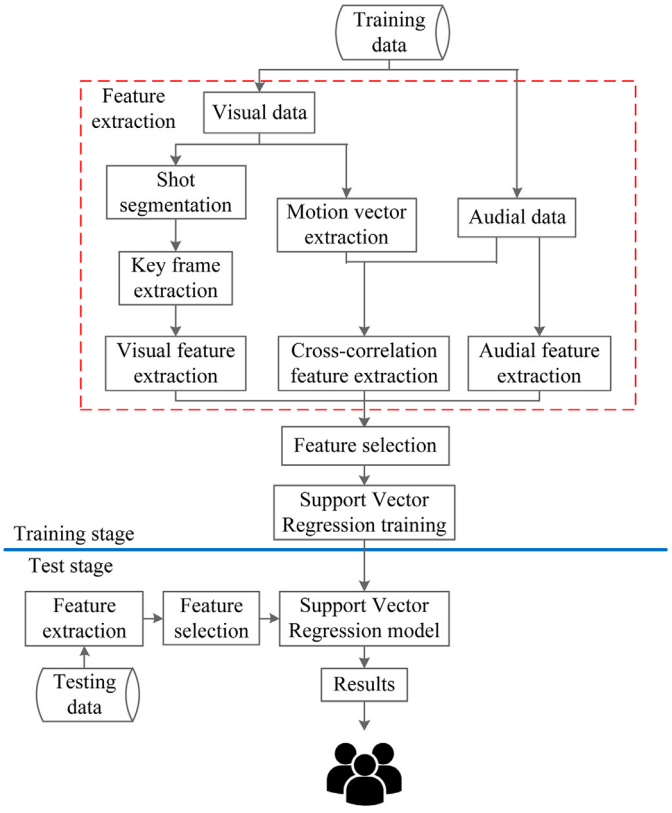

Fig. 6. The framework of the proposed affective video recognition system.

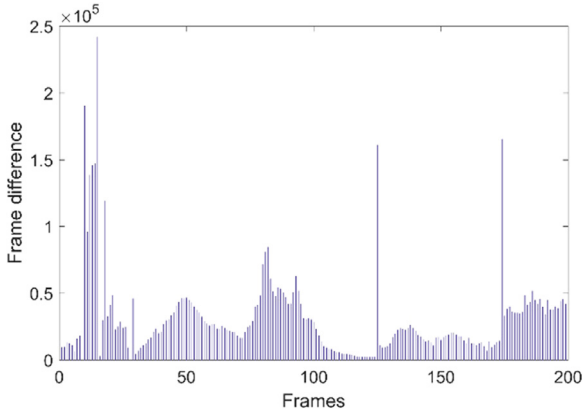

Fig. 7. Shot boundary sequence of a video with the χ^2 histogram algorithm.

4.2. Shot segmentation and key frame extraction for HHT-based visual feature extraction

In order to obtain the HHT-based visual feature, the visual signal is segmented into a series of shots, and the key frame of each shot is extracted. A shot is a basic temporal unit of films and made up of a series of interrelated consecutive frames combined with sounds representing a continuous action in time and space. The aim of shot segmentation is to detect the shot boundary and spilt up a film into shots. In this paper, the χ^2 histogram algorithm in [39] is used for the shot boundary detection and computed as follows:

$$d(I_i, I_j) = \sum_{k=1}^{n} \frac{\left[H_i(k) - H_j(k)\right]^2}{H_i(k)} \tag{19}$$

where $d(I_i, I_j)$ is the distance between two frames; H_i and H_j are the color histograms of the frames I_i and I_j respectively. Fig. 7 is the shot boundary sequence of a video computed with the proposed χ^2 histogram algorithm. After the shot segmentation, the motion attention model is used to extract the key frame of each shot as the best summary of video contents [40].

For each extracted key frame, the two-dimensional HHT algorithm is applied to obtain the HHT-based visual features. It is worthwhile to note that the key frame extraction can reduce the computation complexity of the visual feature extraction.

4.3. Preprocessing for HHT-based audio feature extraction

In contrast to the HHT-based visual features, the HHT-based audio features are directly obtained from the audio signal using the one-dimensional HHT algorithm. For audio signals, the high-frequency components have relatively low amplitude due to radiation effects of videos from clips, which will have a significant impact on the feature extraction at the high end of the spectrum. Therefore, a pre-emphasis is performed before applying the one-dimensional HHT algorithm to audio signals in order to increase the amplitude of the high-frequency components with respect to the magnitude of other frequencies. As a result, the overall signal-to-noise ratio (SNR) is improved by minimizing the radiation effects of the audio signals. The pre-emphasis is computed with a pre-emphasis network:

$$f(x) = 1 - \alpha x^{-1} \tag{20}$$

where α is the emphasized coefficient with a typical value of 0.95. After pre-emphasis, the audio signal can be written as

$$s_{pre}(n) = s(n) - \alpha s(n-1) \tag{21}$$

where $s(n)$ is the nth sample of the input audio signal, and $s_{pre}(0) = s(0)$.

4.4. Motion computation for cross-correlation feature extraction

The cross-correlation feature is a representation of the relationship between the audio and the visual contents of a video clip. In order to obtain the cross-correlation features, an averaged motion vector is first obtained from the original video frames. The motion vectors are computed from neighboring frames and are aimed at representing the relative movement from the reference frame to the target frame, which have been widely used in video encoding/decoding such as H.246/MPEG-4 AVC. The function of the motion vectors in MPEG is to minimize the sum of squared error over a macro-block. Thus, the motion vector is a two-dimensional vector with the same size as the macro-block. For a given frame $\mathbf{mv}(x, y, p)$, the motion vector can be obtained in the decoding process. In order to obtain the motion intensity of video frames, all motion vectors of a frame are normalized as the averaged motion vector with the following equation:

$$m(p) = \frac{\sum_{x=0}^{M_1} \sum_{y=0}^{M_2} ||\mathbf{mv}(x, y, p)||}{\max\{||\mathbf{mv}(x, y, p)||\} M_1 M_2} \tag{22}$$

where M_1 and M_2 are the height and the width of the macro-block, respectively. x and y are the indices of macro-block in the frame p.

Once the averaged motion vector is obtained, the cross-correlation technique is applied to it and to the audio signal. Then the result of the cross-correlation analysis is transformed into the time-frequency domain using the Hilbert transform. Finally, the cross-correlation features are extracted consisting of three components: the instantaneous amplitude, the instantaneous phase and the instantaneous frequency.

4.5. Procedure of the HHTC feature extraction

The following computational steps describe the procedure of HHTC Feature Extraction.

Step 1 Read video clips. For a video clip, the visual and the audio signals are directly read from the files and saved individually.

Step 2 HHT-based visual feature extraction. The visual signal is segmented and a series of shots is obtained for the key frame extraction. The two-dimensional HHT algorithm is then performed on each extracted key frame, and the HHT-based visual features, including the instantaneous amplitude, the instantaneous phase and the instantaneous frequency, are computed with the results of the two-dimensional HHT algorithm.

Step 3 HHT-based audio feature extraction. The HHT-based audio features also include the three components as the HHT-based visual features. A pre-emphasis is firstly performed to improve the overall signal-to-noise ratio before applying the one-dimensional HHT algorithm on the audio signal to obtain the HHT-based audio features.

Step 4 Cross-correlation feature extraction. In order to prepare the visual signal for the cross-correlation analysis, an averaged motion vector is calculated. The cross-correlation technique is then performed on the obtained averaged motion vector combined with the audio signal. The cross-correlation analysis result is calculated using the Hilbert transform and the cross-correlation features, including the instantaneous amplitude, the instantaneous phase, and the instantaneous frequency, are extracted.

4.6. Feature selection

The HHTC features consist of three types and each type has three feature vectors as described in Section 3. Each feature vector has a series of components, of which some are more relevant for a certain goal than others. Thus, a feature selection method, called the Principal Component Analysis (PCA) [41,42], is applied to reduce the number of components of each feature vector. The PCA algorithm is a statistical procedure to convert a set of possibly correlated variables into linearly uncorrelated variables using an orthogonal transformation. The linearly uncorrelated results are called principal components and are less than the original number of variables. Mathematically, the orthogonal transformation in PCA is defined by a set of p-dimensional vectors $\mathbf{W}_{(k)} = (w_1, \ldots, w_p)_{(k)}$ that map the original vector $\mathbf{X}_{(i)}$ to a new vector of principal components $\mathbf{Y}_{(i)} = (y_1, \ldots, y_k)_{(i)}$ as follows:

$$\mathbf{Y}_{ik} = \mathbf{X}_{(i)} \cdot \mathbf{W}_{(k)}. \tag{23}$$

The individual variables of $\mathbf{Y}$ inherit the maximum possible variance from $\mathbf{X}$ with $\mathbf{W}$ constrained to be a unit vector. The PCA transforms the original vector $\mathbf{X}$ to a new coordinate system such that the first coordinate is with the greatest variance, the second coordinate is with the second greatest variance, and so on. The kth component of loading vector $\mathbf{W}$ is computed as follows:

$$\mathbf{W}_{(k)} = \underset{||W=1||}{argmax}\{||\mathbf{X}_{(k)}\mathbf{X}_{(i)}||^2\} \tag{24}$$

where

$$\mathbf{X}_{(i)} = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{W}_{(s)}\mathbf{W}_{(s)}^T. \tag{25}$$

The obtained HHTC features are all combined into a matrix and then PCA is used to map the matrix into a new matrix with fewer components. The PCA result is an uncorrelated orthogonal basis set, and its operation can reveal the internal structure of the data that best matches the variance of the data. The application of a feature selection algorithm allows the learning step to work with significant features and leads to a better understanding of the problem to be solved.

Table 2
Common kernels for SVR.

Kernel	Formula
Linear	$k(x_i, x_j) = x_i * x_j$
Polynomial	$k(x_i, x_j) = (1 + x_i * x_j)^d$
Gaussian radial basis function	$k(x_i, x_j) = \exp(-\gamma\|\|x_i - x_j\|\|^2),\ \gamma > 0$
Hyperbolic tangent	$k(x_i, x_j) = \tanh(\kappa x_i * x_j + c),\ \kappa > 0$ *and* $c < 0$

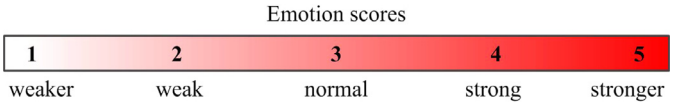

Fig. 8. Five scale of various emotion scores.

4.7. Emotion recognition

Once the selected HHTC features are obtained, a normalization is applied to eliminate the variability of different feature vectors while keeping the emotional discrimination [10]. The extracted features are fed into a nonlinear regression system trained with a high success rate to detect the participants' affective state while watching the videos. The Support Vector Regression (SVR) as a nonlinear regression method is used in this paper for modeling the framework. The SVR is a supervised model with associated learning algorithms and its performance depends on the kernels. Several commonly used kernels for SVR are shown in Table 2.

5. Experimental study

This section presents the experimental results for the proposed approach. First, the two databases used in the experiments are introduced. Then different kernels of SVR for video affective recognition system are compared with each other, and the one with the lowest Root Mean Squared Error (RMSE) is chosen for obtaining the desired output. Finally, the performance of the proposed approach is evaluated on the video datasets with the selected SVR kernel.

5.1. Video dataset

In order to evaluate the performance of the proposed approach, two different datasets are employed for the following experiments. The first dataset named VACAD is conducted by our research team that consists of 112 film videos. These 112 films are manually cut into 2000 video clips representing various genres including drama, comedy, romance, fantasy and sci-fi. The duration of each video clip is around 10 s, and the audio sample rate is 44,100 Hz. Each video clip contains no more than 20 shots. Total 50 participants were asked to rate the 2000 video clips with 5 scores ranging from 1 (weak) to 5 (strong) as shown in Fig. 8. Different participants may have different scores for the same video clip, because the annotation process is subjective. Finally, the annotated results from all participants are averaged and rounded to the nearest integer and are taken as the ground truth of the six basic emotions: anger, fear, happiness, sadness, disgust, and surprise.

The second dataset is the LIRIS-ACCEDE consisting of 9800 video excerpts extracted from 160 movies shared under Creative Commons licenses with a large content diversity. The Russell's valence-arousal scale is used for quantitatively describing the emotions in this database. In this emotion model, each affective state can be described with its position in a two-dimensional space

Table 3
The RMSE of different SVR kernels.

Kernel	RMSE					
	Anger	Fear	Happiness	Sadness	Disgust	Surprise
Linear	0.237	0.185	0.262	0.258	0.192	0.258
Polynomial	0.229	0.221	0.217	0.224	0.202	0.273
Gaussian radial basis function	0.196	0.183	0.203	0.229	0.218	0.221
Hyperbolic tangent	0.211	0.205	0.224	0.275	0.254	0.326

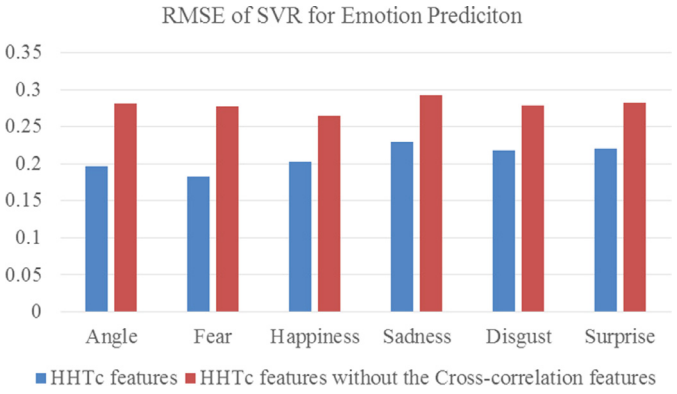

Fig. 9. The RMSE of SVR for emotion prediction with and without the cross-correlation features.

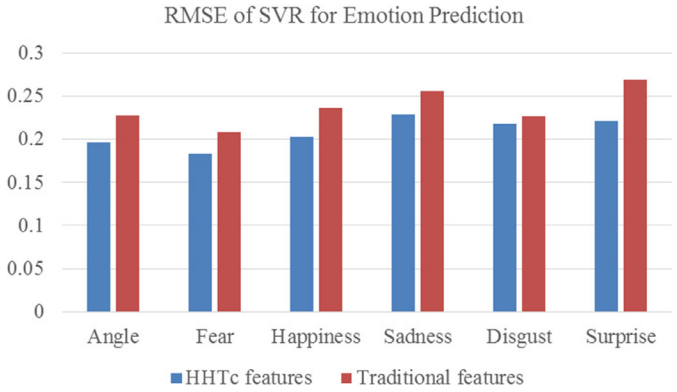

Fig. 10. The RMSE of SVR for emotion prediction with the traditional features and the HHTC features.

formed with valence and arousal. The 9800 video excerpts are sorted along the induced valence axis from the most negatively perceived video to the most positively perceived video. Each video excerpt is 8–12 second long and annotated by 1517 participants from 89 different countries through crowdsourcing. Several movies genres, such as horror, comedy, drama and action etc., are presented in this database.

5.2. Results

In order to evaluate the performance of the proposed HHTC features, several experiments are conducted on the two databases described above. The experimental results on these two databases are presented in the following subsections.

5.2.1. Experimental results with VACAD

A comparison experiment is conducted with the proposed HHTC features and the Ekman's emotion model on the VACAD dataset to find the most appropriate SVR kernel for regression. The kernels used for comparison are presented in Table 2, and the one with lowest RMSE is selected. In this experiment, 1600 of the 2000 video clips are taken as the training data, while the remaining are considered as the testing data. The RMSE of the four kernels are shown in Table 3. It can be seen from the table that the Gaussian radial basis function has the lowest RMSE compared with others under the same experimental conditions.

The proposed HHTC features contain three types of features with each type presenting different video content. The cross-correlation features that present the dependencies between the visual and the audio signals are significant in the HHTC features. An experiment is conducted to validate the importance of the cross-correlation features in video affective content analysis. In this experiment, the Gaussian radial basis function is used as the SVR kernel and the experimental results with and without the cross-correlation features are shown in Fig. 9. It can be clearly seen that the video emotional regression accuracy increases with the cross-correlation features. Thus, the cross-correlation features play an important role in the HHTC features and can help to improve the performance of the video affective content analysis.

The proposed HHTC features are compared with the traditional features on the VACAD dataset in the application of video affective content analysis. The features used in this experiment are presented in Table 4. The first column of Table 4 lists the name of feature set, the second column is the feature type, and the last column presents the features names. The emotion model used in this comparison experiment is described in Section 5.1 that includes six types of emotions. The model for mapping the low-level features to the emotions of videos is the same for all features. Fig. 10 shows the experimental results of SVR for motion prediction based on the proposed HHTC features and the traditional features. The RMSE of SVR based on the proposed HHTC features is lower than the one based on the traditional features for all six emotions. Thus, the proposed features can outperform previous approaches with statistical significance.

Table 4
Features used for comparisons.

Feature set	Type	Features
Traditional features	Visual	lighting, saturation, color energy, shot-change rate, motion intensity, shot duration, shot type transition rate, motion dynamics, color heat
	Audio	zero crossing rate, sound energy, pitch, onset, beat strength, bandwidth, tempo, spectral centroid, spectral rolloff, spectral flux, loudness, MFCC
HHTC features	HHT-based visual HHT-based audio Cross-correlation	instantaneous frequency, instantaneous amplitude, instantaneous phase

Table 5
Features for estimating arousal and valence dimensions of the LIRIS-ACCEDE database [25].

Arousal	Valence
1. Global activity	1. Colorfulness
2. Number of scene cuts per frame	2. Hue count
3. Standard deviation of the wavelet coefficients of audio signal	3. Audio zero-crossing rate
4. Median lightness	4. Entropy complexity
5. Slope of the power spectrum	5. Disparity of most salient points
6. Lighting	6. Audio asymmetry envelop
7. Colorfulness	7. Number of scene cuts per frame
8. Harmonization energy	8. Depth of field
9. Length of scene cuts	9. Compositional balance
10. Audio flatness envelop	10. Audio flatness

Table 6
Regression results with different feature sets on the LIRIS-ACCEDE database.

Feature set	Arousal		Valence	
	MSE	Pearson's r	MSE	Pearson's r
Features from Table 5	0.303	0.308	0.302	0.310
HHTC features	0.294	0.321	0.290	0.327

5.2.2. Experimental results with LIRIS-ACCEDE

A comparison experiment is conducted on the LIRIS-ACCEDE database to further validate the proposed approach. Two independent SVRs are used in this experiment to separately model the arousal and the valence. The Gaussian radial basis function is also selected as the kernel function. The ground truth is made up of two subsets: the lowest valence and the highest valence, which is achieved by binarizing the database. The training and the test data each contain 4900 video clips, thus the distribution of the movies genre in the training and the test data is the same. The features used for estimating the arousal and the valence dimensions of the LIRIS-ACCEDE database are shown in Table 5. All features used in this comparison experiment are normalized using the Z-score algorithm before the learning step.

The experimental results are shown in Table 6. The Mean Squared Error (MSE) and the Pearson correlation coefficient (Pearson's r) are computed to evaluate the performance of emotion regression. The MSE is widely used for regression models to measure the errors between the estimated and the expected values, which means that the amount of the estimated values differing from the ground truth is measured. The Pearsons r is a measure of linear correlation between two values. It can be seen from Table 6 that the HHTC features have the lowest MSE and the highest absolute value of Pearson's r. Thus, the proposed approach can achieve a better overall performance with different emotion models and databases.

5.2.3. Computation analysis

Before obtaining the HHT-based visual features, a shot segmentation is applied on the visual signal, and then the key frame of each shot is extracted. Generally, the duration of a shot is about 3–5 s, and there are at least 24 frames per second. Based on these conditions, the HHT-based visual feature extraction does not need to deal with every frame but only the key frame. The information between the adjacent frames is redundant within the same shot, and mostly the key frame can represent all the information of a shot. Thus, the computational load will be reduced by more than 70 times compared with the traditional feature extraction processes.

6. Conclusion

In this paper, a video affective content analysis technique is presented for social media video analysis. The proposed approach is based on the HHT and the cross-correlation technique. It is used for affective feature extraction in video affective recognition system. The extracted features called the HHTC features consist of three feature types and cover the audio content, the visual content, and the dependencies between the audio and the visual signals. Each type of the HHTC features has three components: the instantaneous frequency, the instantaneous amplitude and the instantaneous phase. The learning model based on the SVR and the Ekman's emotion model is used to bridge the gap between the proposed affective features and the affective states of the users. The experiment results have demonstrated and validated that the proposed method can effectively improve the performance of emotion recognition. In addition, the comparison results have also proved that the proposed HHTC features have better discriminative power with the operation on videos compared with the traditional features.

Acknowledgment

We would like to extend our heartfelt thanks to all the reviewers, without whose assistance the accomplishment of this thesis would have been impossible.This work was supported by the National Natural Science Foundation of China (61572060, 61772060) and CERNET Innovation Project (NGII20151004, NGII20160316).

References

[1] G.M. Smith, Film Structure and the Emotion System, Cambridge University Press, 2003, doi:10.1017/CBO9780511497759.
[2] C. Dorai, S. Venkatesh, Computational media aesthetics: Finding meaning beautiful, Fems Microbiol. Lett. 8 (4) (2001) 10–12, doi:10.1007/978-1-4899-7993-3_1036-2.
[3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125, doi:10.1016/j.inffus.2017.02.003.
[4] A. Hanjalic, Adaptive extraction of highlights from a sport video based on excitement modeling, IEEE Trans. Multim. 7 (6) (2005) 1114–1122, doi:10.1109/TMM.2005.858397.
[5] A. Hanjalic, L.Q. Xu, Affective video content representation and modeling, IEEE Trans. Multim. 7 (1) (2005) 143–154, doi:10.1109/TMM.2004.840618.
[6] A. Hanjalic, Extracting moods from pictures and sounds: towards truly personalized TV, IEEE Signal Process. Mag. 23 (2) (2006) 90–100, doi:10.1109/MSP.2006.1621452.
[7] T. Eerola, J.K. Vuoskoski, A comparison of the discrete and dimensional models of emotion in music, Psychol. Music 39 (1) (2011) 18–49, doi:10.1177/0305735610362821.
[8] I.C. Christie, B.H. Friedman, Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach, Int. J. Psychophysiol. 51 (2) (2004) 143, doi:10.1016/j.ijpsycho.2003.08.002.
[9] P. Ekman, An argument for basic emotions, Cognit. Emot. 6 (3–4) (1992) 169–200, doi:10.1080/02699939208411068.
[10] J.M. Ren, M.J. Wu, J.S. Jang, Automatic music mood classification based on timbre and modulation features, IEEE Trans. Affect. Comput. 6 (3) (2015) 236–246, doi:10.1109/TAFFC.2015.2427836.
[11] H.L. Wang, L.F. Cheong, Affective understanding in film, IEEE Trans. Circuits Syst. Video Technol. 16 (6) (2006) 689–704, doi:10.1109/TCSVT.2006.873781.
[12] S. Zhang, Q. Huang, S. Jiang, W. Gao, Q. Tian, Affective visualization and retrieval for music video, IEEE Trans. Multim. 12 (6) (2010) 510–522, doi:10.1109/TMM.2010.2059634.
[13] S. Zhang, Q. Tian, S. Jiang, Q. Huang, Affective MTV analysis based on arousal and valence features, in: IEEE International Conference on Multimedia and Expo, (ICME 2008, June 23–26 2008), Hannover, Germany, 2008, pp. 1369–1372, doi:10.1109/ICME.2008.4607698.
[14] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (1980) 1161–1178, doi:10.1037/h0077714.
[15] Y. Cui, J.S. Jin, S. Zhang, S. Luo, Q. Tian, Music video affective understanding using feature importance analysis, in: ACM International Conference on Image and Video Retrieval, Civr 2010, Xi'an, China, July, 2010, pp. 213–219, doi:10.1145/1816041.1816074.
[16] J. Niu, X. Zhao, L. Zhu, H. Li, Affivir: an affect-based internet video recommendation system, Neurocomputing 120 (10) (2013) 422–433, doi:10.1016/j.neucom.2012.07.050.

[17] N.E. Huang, Z. Wu, A review on Hilbert-Huang transform: method and its applications to geophysical studies, Rev. Geophys. 46 (2) (2008) RG2006, doi:10.1029/2007RG000228.
[18] J. Ortega, G.H. Smith, Hilbert–Huang transform analysis of storm waves, Appl. Ocean Res. 31 (3) (2009) 212–219, doi:10.1016/j.apor.2009.09.003.
[19] C. Shahnaz, S. Sultana, S.A. Fattah, R.H.M. Rafi, Emotion recognition based on EMD-Wavelet analysis of speech signals, in: IEEE International Conference on Digital Signal Processing, ICDSP 2015, 21–24 July 2015, 2015, pp. 307–310, doi:10.1109/ICDSP.2015.7251881.
[20] A. Kotowski, Frequency analysis with cross-correlation envelope approach, Acta Mech. Autom. 6 (4) (2012) 27–31.
[21] I. Noda, Determination of two-dimensional correlation spectra using the Hilbert transform, Appl. Spectrosc. 54 (7) (2000) 994–999, doi:10.1366/0003702001950472.
[22] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 28 (7) (2008) 779–784.
[23] R.G. Brereton, G.R. Lloyd, Support vector machines for classification and regression, Analyst 135 (2) (1998) 230, doi:10.1039/B918972F.
[24] B. Yoann, B.J. Noel, D. Emmanuel, C. Liming, C. Christel, A large video database for computational models of induced emotion, in: IEEE Affective Computing and Intelligent Interaction, 7971, 2013, pp. 13–18, doi:10.1109/ACII.2013.9.
[25] B. Yoann, D. Emmanuel, C. Christel, C. Liming, LIRIS-ACCEDE: a video database for affective content analysis, IEEE Trans. Affect. Comput. 6 (1) (2015) 43–55, doi:10.1109/TAFFC.2015.2396531.
[26] M.M. Bradley, M.K. Greenwald, M.C. Petry, P.J. Lang, Remembering pictures: pleasure and arousal in memory, J. Exp. Psychol. Learn. Memory & Cognit. 18 (2) (1992) 379, doi:10.1037/0278-7393.18.2.379.
[27] D. Watson, A. Tellegen, Toward a consensual structure of mood, Psychol. Bull. 98 (2) (1985) 219–235, doi:10.1037/0033-2909.98.2.219.
[28] E. Cambria, A. Livingstone, A. Hussain, The hourglass of emotions, in: International Conference on Cognitive Behavioural Systems, COST 11, 7403, 2011, pp. 144–157, doi:10.1007/978-3-642-34584-5_11.
[29] A. Mehrabian, Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies, Cambridge, MA: Oelgeschlager, 1980.
[30] H. Lövheim, A new three-dimensional model for emotions and monoamine neurotransmitters, Med. Hypotheses 78 (2) (2012) 341–348, doi:10.1016/j.mehy.2011.11.016.
[31] K. Hevner, Expression in music: a discussion of experimental studies and theories, Psychol. Rev. 42 (2) (1935) 186–204, doi:10.1037/h0054832.
[32] B.S. Canini Luca, L. Riccardo, Affective recommendation of movies based on selected connotative features, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 636–647, doi:10.1109/TCSVT.2012.2211935.
[33] A. Yazdani, K. Kappeler, T. Ebrahimi, Affective content analysis of music video clips, in: International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies, NY, USA, 2011, pp. 7–12, doi:10.1145/2072529.2072532.
[34] S. Wang, Q. Ji, Video affective content analysis: a survey of state of the art methods, IEEE Trans. Affect. Comput. 6 (4) (2015) 410–430, doi:10.1109/TAFFC.2015.2432791.
[35] N.E. Huang, S.R. Long, Z. Shen, The mechanism for frequency downshift in nonlinear wave evolution, Adv. Appl. Mech. 32 (8) (1996) 59–117, doi:10.1016/S0065-2156(08)70076-0.
[36] F. King, Hilbert Transforms, Cambridge University Press, Cambridge U.K., 2009, doi:10.1017/CBO9780511735271.
[37] M.S. Koh, E. Rodriguez-Marek, Perfect reconstructable decimated two-dimensional empirical mode decomposition filter banks, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 26–31 May 2013, 2013, pp. 1952–1956, doi:10.1109/ICASSP.2013.6637994.
[38] N.E. Huang, Z. Shen, S.R. Long, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. A Math. Phys. Eng. Sci. 454 (1971) (1998) 903–995, doi:10.1098/rspa.1998.0193.
[39] H. Jiang, A. Helal, A.K. Elmagarmid, A. Joshi, Scene change detection techniques for video database systems, Multim. Syst. 6 (3) (1998) 186–195 doi:1007/s005300050087.
[40] C. Sujatha, U. Mudenagudi, A study on keyframe extraction methods for video summary, in: International Conference on Computational Intelligence and Communication Networks, 2011, pp. 73–77, doi:10.1109/CICN.2011.15.
[41] K. Pearson, On lines and planes of closest fit to systems of points in space, Lond., Edinb. Dublin Philos. Mag. J. Sci. (1901) 559–572, doi:10.1080/14786440109462720.
[42] H.H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educat. Psychol. 24 (6) (2010) 417–441, doi:10.1037/h0071325.

Shasha Mo received the B.S. degree in School of Electronic and Information Engineering from Beijing Jiaotong University, Beijing, in 2010 and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2015. She is currently a postdoctoral researcher with the School of Computer Science and Engineering, Beihang University. Her current research interests include affective computing and music mood analysis.

Jianwei Niu received the Ph.D. degree in computer science from Beijing University of Aeronautics and Astronautics (BUAA, now Beihang University), China, in 2002. He was a visiting scholar at School of Computer Science, Carnegie Mellon University, USA, from Jan. 2010 to Feb. 2011. He is a professor in the School of Computer Science and Engineering, BUAA. He has published more than 100 referred papers, and filed more than 30 patents in mobile and pervasive computing. He served as the Program Chair of IEEE SEC 2008, Executive Co-chair of TPC of CPSCom 2013, TPC members of InfoCom,Percom, ICC, WCNC, Globecom, LCN, and etc. He has served as associate editor of International Journal of Ad Hoc and Ubiquitous Computing, associate editor of Journal of Internet Technology, editor of Journal of Network and Computer Applications (Elsevier). He received the New Century Excellent Researcher Award from Ministry of Education of China 2009, the first prize of technical invention of the Ministry of Education of China 2012, Innovation Award from Nokia Research Center, and won the best paper award in IEEE ICC 2013, WCNC 2013, ICACT 2013, CWSN 2012 and GreenCom 2010. His current research interests include affective computing, mobile and pervasive computing, mobile video analysis. He is a senior member of the IEEE.

Yiming Su received the B.S. degree from School of Computer Science and Engineering, Beihang University, Beijing, in 2014 and the M.S. degree from School of Computer Science and Engineering, Beihang University, Beijing, in 2017. His current research interests include affective computing and video emotion analysis.

Sajal K. Das is currently the chair at Computer Science Department and Daniel St. Clair Endowed chair at the Missouri University of Science and Technology. Prior to that he was a distinguished scholar professor of computer science and engineering and the founding director of the Center for Research in Wireless Mobility and Networking (CReWMaN) at the University of Texas at Arlington (UTA). His current research interests include wireless and sensor networks, mobile and pervasive computing, cyberphysical security, distributed and cloud computing, biological and social networks, applied graph theory and game theory. He has published more than 500 technical papers and 47 invited book chapters in these areas, holds five US patents, and received nine Best Paper Awards in international conferences. He is a senior member of the IEEE.