Video affective content analysis: a survey of state-of-the-art methods

Shangfei Wang, *Member, IEEE,* Qiang Ji, *Fellow, IEEE,*

Abstract—Video affective content analysis has been an active research area in recent decades, since emotion is an important component in the classification and retrieval of videos. Video affective content analysis can be divided into two approaches: direct and implicit. Direct approaches infer the affective content of videos directly from related audiovisual features. Implicit approaches, on the other hand, detect affective content from videos based on an automatic analysis of a user's spontaneous response while consuming the videos. This paper first proposes a general framework for video affective content analysis, which includes video content, emotional descriptors, and users' spontaneous nonverbal responses, as well as the relationships between the three. Then, we survey current research in both direct and implicit video affective content analysis, **with a focus on direct video affective content analysis.** Lastly, we identify several challenges in this field and put forward recommendations for future research.

Index Terms—Video affective content analysis, emotion recognition, and content-based video retrieval.

1 Introduction

Recent years have seen a rapid increase in the size of video collections. Automatic video content analysis are needed in order to effectively organize video collections and assist users in quickly finding videos. Historically, videos were primarily used for **entertainment** and information-seeking. Thus, conventional content-based video analysis focuses on generic semantic content, such as news or sports. As online services like YouTube and Tudou grow, videos have become the medium for many people to communicate and to find entertainment. These growing video collections will inevitably influence users'emotional states as they spread information and provide entertainment. More and more people watch videos in order to satisfy certain emotional needs (e.g., relieve boredom); it is therefore necessary to tag videos based on their affective content. Unlike conventional content-based video analysis, which typically identifies the main event involved in a video, video affective content analysis is to identify videos that can evoke certain emotions in the users. Introducing such a personal or human touch into video content analysis is expected to benefit both the users and businesses that create, distribute, and host the videos. For example, movie directors may adapt their editing to optimize the emotional flow with the help of the detected audiences' emotional states. Users could retrieve certain videos by inputting their emotional demands. Distributors could select the best target population for their videos based on the affective content.

As a result of these developments, video affective content analysis is becoming increasingly important. The goal of video affective content analysis is to automatically tag each video clip by its affective content. Due to the difficulty in defining objective methods to automatically assess the emotions of a video, the research topic of video affective content analysis has not been thoroughly explored until recently. In 2001, Nack et al. [1] defined the concept of Computational Media Aesthetics (CMA) as the algorithmic study to analyze and interpret how the visual and aural elements in media evoke audiences' emotional responses based on the film grammar. The core trait of CMA is to interpret the data with the makers' eyes. Based on the expected mood, i.e., the emotions a film - maker intends to communicate to a particular audience with a common cultural background, Hanjalic and Xu [2] successfully related audiovisual features with the emotional dimension of the audience. Earlier research has attempted to infer the affective content of videos directly from the related audiovisual features. This kind of research represents the mainstream research on video affective content analysis and is referred to as direct video affective content analysis.

In addition to direct approaches, recent research on video affective content analysis includes inferring the video's affective content indirectly based on an analysis of a user's spontaneous reactions while watching the video. We refer to this kind of research as implicit video tagging. Figure 1 summarizes the major components of the two approaches for video affective content analysis.

Video affective content analysis consists of video content, users' spontaneous nonverbal responses, emotional descriptors, and their relationships [4]. The video content is represented by various visual and audio features. Emotional descriptors capture the users' subjective evaluation of the videos' affective content. Two kinds of descriptors are often used: the categorical approach and the dimensional approach. The users' spontaneous nonverbal responses in-

Shangfei Wang is with the Key Lab of Computing and Communication Software of Anhui Province, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, P.R.China, 230027. E-mail:sfwang@ustc.edu.cn.
Qiang Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA, 12180.. E-mail:qji@ecse.rpi.edu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

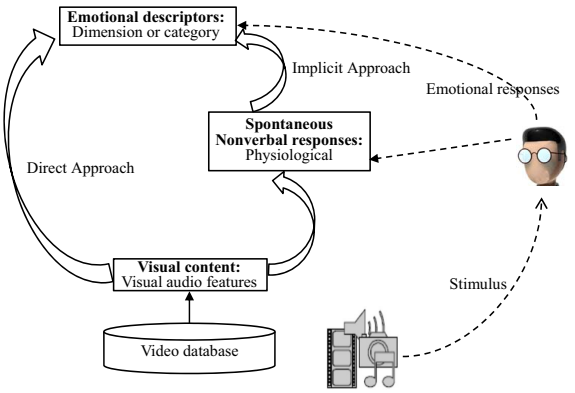

Fig. 1: Components and two major approaches of video affective content analysis [3]

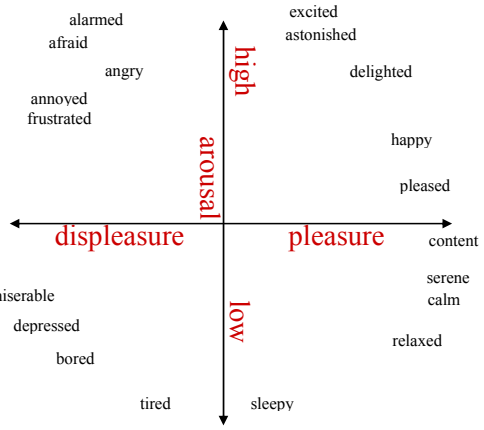

Fig. 2: Mapping of the categorical emotions in valence-arousal space [36]

clude users' physiological and visual behavioral responses while watching videos. The mapping from video content space to emotional descriptor space can be regarded as direct video affective content analysis, while the mapping from users' spontaneous nonverbal response space to emotional descriptor space takes an implicit approach to video affective content analysis. 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 descriptors.

In the sections to follow, we first discuss emotional descriptors. We then review current techniques for direct and implicit video affective content analysis, as well as the existing benchmark datasets for video affective content analysis. The paper concludes with a discussion of challenges and future research directions.

To the best of our knowledge, this work is the first paper to provide a comprehensive review of video affective content analysis. Previous reviews of this topic include the recent work by Soleymani and Pantic [5]. This work, however, only provides a review of implicit tagging, while this paper covers both implicit and direct approaches.

2 Emotional Descriptors

Psychologists have used two major methods to measure emotion: the discrete approach and the dimensional approach. According to Ekman [6], emotion can be grouped into six different categories such as happiness, sadness, surprise, disgust, anger, and fear. First introduced by Wundt [7], the dimensional approach divides emotion into 3D continuous spaces: arousal, valence, and dominance. Based on these theories, two kinds of emotional descriptors have been proposed to capture a video's affective content. One is the categorical approach, and the other is the dimensional approach.

Many emotion theorists have claimed that there is a set of basic emotional categories. However, there is some argument over which emotions belong in this basic set. The most frequently used categories in the field of video affective content analysis are Ekman's six basic emotions [6], including happiness, sadness, anger, disgust, fear, and surprise [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. In addition, other categories, such as amusement [18], [19], boredom [20], excitement [21], [22], or horror [23], [24], [25], [26], [20] are also used to describe affective content of videos for certain applications.

Dimensional views of emotion have been advocated and applied by a several researchers. Most agree that three dimensions are enough to describe a subjective response. However, a consensus has not been reached on the labels of the dimensions. Valence-arousal-dominance is one set of labels. Arousal measures the activation level of the emotion. As a measure of excitement, arousal characterizes a state of heightened physiological activity and ranges from passive to active or excited. Valence measures the degree of pleasure. It represents a "good feeling" or a "bad feeling", and it ranges from pleasant to unpleasant. Dominance represents the controlling and dominant nature of the emotion. It ranges from submissive (or without control) to dominant (or in control/empowered). Dominance is difficult to measure and is often omitted, leading to the commonly used two dimensional approach. Instead of valence-arousal-dominance, a few work adopt other dimensions. For example, Canini et al. [27] proposed to use natural, temporal, and energetic dimensions. Some works [28], [29] discretize dimension description into categories, such as positive and negative valence and high and low arousal. Others [30], [2], [31], [32], [33], [34], [35] use continuous dimensional descriptors.

The categorical and dimensional definitions of emotion are related. In fact, the categorical emotional states can be mapped into the dimensional space. For example, a relaxed state relates to low arousal, while anger relates to high arousal. Positive valence relates to a happy state, while negative valence relates to depressed or angry state. Anger is a dominant emotion, while fear is a submissive emotion. Figure 2 shows the emotional circumplex [36] that plots the categorical emotions in the valence-arousal (VA) space.

Both discrete and continuous descriptors have their limits. Since emotions are complex and subjective, a small number of discrete categories may not reflect the subtlety and complexity of the affective states [37]. A continuous emotional space may embody more subtle and fuzzy e-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

motions without boundaries. It also avoids the problem of predefining a number of emotion categories. However, absolute continuous scores may not be meaningful due to lack of agreed-upon standards for subjective emotion rating.

3 Direct video affective content analysis

As an emerging area of research, video affective content analysis is attracting increasing attention from different fields ranging from multimedia, psychology, entertainment to computer vision. The first research in direct video affective content analysis dates back to 2005, when Hanjalic and Xu [2] proposed a computational framework for affective video modeling and classification in the arousal and valence space. Since then, much work has been done in this field.

The framework of direct video affective content analysis mainly consists of two parts: video feature extraction and emotion classification/regression. First, several visual and audio features are extracted from videos to characterize the video content. Then, the extracted features are fed into a general purpose classifier or regressor such as Support Vector Machine (SVM) [15], [23], [12] or Support Vector Regression(SVR) [27], [30], [38], [39], [31] for emotion classification or regression. Table 1 and Table 2 provide an exhaustive summary of direct video affective content analysis works that respectively use continuous emotion dimensions or discrete emotion categories, as well as extracted features, adopted classifiers/regressors, emotion descriptors, size of the dataset **(i.e., the number of video clips if not explicitly stated)**, number of annotators, and experimental results. Below, we review the audio and visual features used to capture a video's affective content, as well as classification methods that map video features to a video's emotional descriptors.

3.1 Affective feature extraction

The video content can be captured by various visual and audio features. Specifically, the affective content of a video consists of two main categories of data: visual data and auditory data. The visual data can be further divided into visual image, print, and other graphics, while the auditory signal can be divided into speech, music, and environmental sound. An important issue in video affective content analysis is the extraction of suitable low-level acoustic and visual features that effectively characterize different affective content. Both cinematography and psychological research show that certain audio-visual cues are related to the affective content of a video [15]. As a result, cinematic principles and psychological findings have played an important role in defining visual and audio features for characterizing a video's affective content. In the sections to follow, we review audio and visual features used for video affective content characterization, with an emphasis on cinematic and psychological principles.

3.1.1 Audio features

Audio features are essential in characterizing a video's affective content. In fact, Wang and Cheong's study [15] shows that audio features are often more informative than visual ones with respect to affective content characterization.

The first step in acoustic feature extraction is audio type segmentation (also called audio source separation), since the audio part of a video often contains a mixture of sounds from different sources. Audio type segmentation divides the audio part of a video into speech, music, and environmental sound. Wang and Cheong [15] used two features (chroma difference and low short time energy ratio) to distinguish music sound from environmental sound with a simple SVM for every two-second segment of audio signal. Outside the field of video affective content analysis, there is research focusing specifically on audio type segmentation beyond video affective content analysis. Lu et al. [51] introduced a technique to segment and classify audio signals into speech, music, and environmental sounds. Their method first segments a signal into speech and non-speech using such features as high zero crossing rate ratio, low short time energy ratio, linear spectral pairs, and spectrum flux. The non-speech signal is then further divided into music and environmental sound using band periodicity, noise frame ratio, and spectrum flux. Bachu et al. [52] proposed to use zero-crossing rate and energy features to separate speech and non-speech signals. More recently, Radmard et al. [53] proposed a clustering method to separate speech and non-speech signals based on the analysis of cepstral peak, zero-crossing rate, and autocorrelation function peak of short time segments of the speech signal. Zhang and Kuo [54] proposed to use Zero Crossing Rate (ZCR) to separate audio signals into music, speech, and environmental sounds.

Given the segmented audio signal, acoustic features can then be extracted separately from speech, music, and environmental sound. Due to the predominance of speech in characterizing a video's emotion as well as extensive research in speech emotion recognition, much work has been invested in speech feature extraction.

For the speech channel, acoustic feature extraction efforts largely follow the work in speech emotion recognition, which is an active and well-studied field. Psychological research has shown that psycho-physiological characteristics like air intake, vocal muscle, intonation and pitch characteristics vary with emotions [55]. Based on such physiological studies, prosodic speech features are typically used to characterize the affective content of an utterance, since prosody captures the emotional state of the speaker through its non-lexical elements including the rhythm, stress, and intonation of speech [22]. Typical prosodic speech features include loudness, speech rate, pitch, inflection, rhythm, and voice quality. Some speech features are more informative in capturing speech emotions. For example, a study by Scherer and Zentner [56] shows that tempo and loudness are powerful signals of emotional meaning. Xu et al. [48] show that pitch is significant for emotion detection, especially for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

TABLE 1: Direct video affective content analysis using continuous emotion dimension

References	Features	Regressors	Emotion descriptors	♯**Video clips**	♯**Annotators**	Results
Canini et al. [27]	A: sound energy, low-energy ratio, ZCR, spectral rolloff, spectral centroid, spectral flux, MFCC, subband distribution, beat histogram, rhythmic strength, V: color, lighting key, saturation, motion, shot length,shot type transition rate	SVR	natural, temporal,and energetic dimension	75	almost 300	the Kendall's tau metric, $K_{\Delta c} = 0.425$, $K_{\Delta \frown C} = 0.467$, $K_{\Delta \frown w} = 0.502$
Cui et al. [30]	A: ZCR, short time energy, pitch, bandwidth, brightness, roll off, spectrum flux, sub-band peak, sub-band valley, sub-band contrast,tempo, rhythm strength, rhythm contrast, rhythm regularity, onset frequency, drum amplitude V: motion intensity, shot switch rate, frame predictability, frame brightness, frame saturation, colour energy	SVR	arousal and valence	655	N/A	Variance of absolute error , Arousal: 0.107, Valence: 0.118
Cui et al. [38]	A: pitch, tempo, rhythm regularity, sound energy, ZCR, beat strength, bandwidth, rhythm strength, short time energy, V: shot switch rate, lighting, saturation, color energy, motion intensity	SVR and MLR	arousal and valence	552	11	Mean absolute error, Arousal: 0.340, Valence: 0.277
Hanjalic et al. [40], [2], [41]	AR: motion activity, density of cuts, sound energy VA: pitch.	Defined curves	arousal and valence	2	N/A	The proposed arousal and valence models can represent affective video content.
Zhang et al. [39]	AR: ZCR, short time energy, sub-band Peak, sub-band Valley, sub-band Contrast, tempo, rhythm strength, rhythm contrast, rhythm regularity, drum amplitude, motion intensity, short switch rate, frame brightness. VA: pitch, sub-band peak, sub-band valley, sub-band contrast, pitch STD, rhythm regularity, frame brightness, saturation, color energy	SVR	arousal and valence	4000	N/A	P for arousal: 0.620, P for valence: 0.580
Zhang et al. [31]	AR: motion intensity, short switch rate, zero crossing rate, tempo, and beat strength. VA: lighting, saturation, color energy, rhythm regularity, and pitch	SVR	arousal and valence	552	37 (9F,28M)	R: 0.684, P: 0.701
Zhang et al. [42]	AR: motion intensity, shot switch rate, sound energy, zero crossing rate, tempo and beat Strength, VA: rhythm regularity, pitch, lighting, saturation and color energy	affinity propagation	arousal and valence	156	11 (1F,10M)	Arousal P: 0.929

A: Audio features; V: Visual features; AR: Arousal features; VA: Valence features; P: Precision; R: Recall; F: Female; M: male.

the emotions in speech and music. Sound energy is also an important speech feature [2], [27].

Further research has shown that certain speech features are good at capturing certain types of emotions. Continuous features such as speech energy, pitch, timing, voice quality, duration, fundamental frequency, and formant are effective in classifying high or low arousal. This is supported by findings in [57] and [2], which show that inflection, rhythm, voice quality, and pitch are commonly related to valence, while loudness (i.e. speech energy) and speech rate relate to arousal. For discrete emotions, features like pitch levels can indicate feelings such as astonishment, boredom, or puzzlement, while speech volume is generally representative of emotions such as fear or anger. In addition to time-domain features, certain spectral speech features are found to be effective in characterizing speech emotions, since the speech energy distribution varies with emotion. For example, happy speech has a high energy at high frequency range, while sad speech has low energy at the same frequency range [58]. A study in [59] shows that spectral features, like Mel-frequency Cepstrum Coefficients (MFCC), are the most effective features at estimating the continuous values of emotions in 3D space. For a comprehensive review of speech features for speech emotion recognition, readers are advised to refer to [60].

Similar speech features have been employed for video affective content analysis. Wang and Cheong [15] extracted 12 audio features to capture a film's affective content, including energy statistics, Log Frequency Power Coefficients (LFPC), MFCC, and zero-crossing rate statistics. Canini et al. [27] proposed to use low energy ratio and zero crossing rate to capture sound energy, and to use spectral rolloff, spectral centroid, spectral flux, and MFCC to capture the spectral properties of the speech. Xu et al [48] employed arousal-related features including short energy features and MFCC, as well as valence-related audio features, such as pitch. Xu et al. [61] used short time energy for vocal emotion detection, and MFCC to detect excited and non-excited moments. Teixera et al. [11] used zero crossing rate, the irregularity of the spectrum, spectral rolloff, MFCC, etc.

Research on emotion recognition from music [62], [63] has demonstrated that certain aspects such as music mode, intensity, timbre, and rhythm are important in characterizing musical emotions. Common acoustic features for music emotion classification are dynamics (i.e., Root-Mean-Square (RMS) energy), timbre (i.e., MFCCs, spectral shape, spectral contrast), harmony (i.e., roughness, harmonic change, key clarity, and majorness), register (i.e., chromagram, chroma centroid, and deviation), rhythm (i.e., rhythm strength, regularity, tempo, beat histograms), and articulations (i.e., event density, attack slope, and attack time) [64]. Zhang et al. [31] employed rhythm-based music-related features including tempo, beat strength, and rhythm regularity to analyze affective content of music videos. Yang and Chen [65] as well as Eerola and Vuoskoski [66] summarized recent research of emotion and music from the perspective of both informatics and psychology.

For the environment sound channel, certain sound patterns are often used to induce certain emotions. For example, Moncrieff et al. [47] used the changes in sound energy intensity to detect the four sound energy events, i.e., surprise or alarm, apprehension, surprise followed by sustained alarm, and apprehension building up to a climax. The four sound energy events are further used to distinguish between horror and non-horror movies.

In summary, acoustic features are extracted from speech,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

TABLE 2: Direct video affective content analysis using discrete emotion categories

References	Features	Classifiers	Emotion descriptors	♯**Video clips**	♯**Annotators**	Results
Arifin et al. [18], [19]	A: tempo histogram, daubechies wavelet cumulative histogram, MFCC, root mean square energy, spectral flux, spectral rolloff, ZCR V: color, salience and visual tempo	4-level DBNs	sadness, violence, neutral, fear, happiness and amusement	34	14	Accuracy: 0.860
Canini et al. [43]	A: audio track energy V: light source colour, motion dynamics	Define natural, temporal and energetic curves	warm, cold, dynamic, slow, energetic and minimal	87	the first two genres given by Internet Movie Database	the combination of the three emotional dimensions is better than one dimension in terms of P-R curves.
Chan et al. [44], [45]	AR: global motion, shot cut rate, audio energy; VA: colour brightness, colour saturation, pitch;	defined affect curves; Okapi BM25 model	151 emotional labels	39	8	R: 0.800.
Ding et al. [24]	A: MFCCs, spectral power, spectral centroid V: emotional intensity, color harmony, variance of color, lighting key, texture	Multi-view MIL	horror, non-horror	800	N/A	F-measure: 0.843
French [46]	A: sound energy; V: object motion of the main object or character.	define the relation between slapstick and features	slapstick or no slapstick	16	3	R: 0.750; P: 1.000;
Irie et al. [14]	A: pitch, short-term energy, MFCCs V: gravity centers of color, brightness of image, motion intensity, shot duration	LDA	joy, acceptance, fear, surprise, sadness, disgust, anger and anticipation	206	16	subject agreement rate: 0.855
Kang [8]	V: color, motion, shot cut rate	HMMs	fear, sadness and joy	6	10	Accuracy: 0.876
Kang [9], [8]	V: color, motion, and shot cut rate	AdaBoost + relevance feedback	fear, sadness and joy	10	10	Accuracy: 0.849
Moncrieff et al. [47]	A: the dynamics of the sound energy of the audio	define rules between four sound energy events and features	surprise, apprehension or the emphasis of an event, surprise (followed by sustained alarm), builds apprehension to a climax	4	N/A	Positive Support: 0.680, 0.890, 0.930, 0.800
Sun et al. [10]	AR: motion intensity, shot cut rate and sound energy; A: speech rate, pitch average, pitch range, sound energy, silence ratio; V: camera motion(phase/intensity), shot cut rate, color features	defined excitement curves + HMM	joy, anger, sadness and fear	557	30	R: 0.809, P: 0.874, F-Score: 0.840
Teixeira et al. [11]	A: ZCR, irregularity of the spectrum, spectral rolloff, tristimulus features, fundamental frequency, MFCC, root mean square (RMS), energy and the temporal centroid; V: shot length, motion history image, lighting key, color activity, color heat and color weight;	HMM	sadness, happiness, anger, fear, disgust and surprise	346	16 (8F, 8M)	Accuracy: 0.770
Wang et al. [15]	A: energy; LFPC and its delta; low energy ratio; spectral roll-off and centroid; MFCC and its delta; ZCR; spectral flux and normalized version; chroma and its delta; normalized chroma; LSTER; bands statistics; music scale; V: Shot Duration, lighting key, motion, shot density, color energy, miscellaneous.	SVMs	anger, sad, fear, joyous, surprise, tender, and neutral	36	3	Accuracy: 0.858.
Watanapa et al. [22]	A:the average volume difference; V: average squared motion magnitude, average shot duration, average percent of pixels in the clip with high brightness, with dark value, and with cold hue;	multi-sieving NN	excitement, joy and sadness	120	30 (14F,16M)	Accuracy: 0.978
Xu et al. [23]	Case 1: A: MFCC, energy, delta and acceleration, amplitude change of audio signal Case 2: A: energy, pitch, MFCC and LPCC Case 3: affective script partitions; A: MFCC, energy, delta and acceleration	1: SVM + HMM 2: defined relations between script and potential affective partition + HMM 3: defined relations between AEE and affective contents	1: horror and non-horror 2: laughable and non-laughable 3: 3 intensity levels for anger, sadness, fear, joy, love	1: 80 minutes 2: 4 hour segments 3:560 minutes	1: 5 2: 10 (5F, 5M) 3: 8	1: R: 0.976 P: 0.913 2: R: 0.873, P: 0.884 3: R: 1.000, P: 1.000
Xu et al. [12]	A: short-time energy, MFCC, pitch; V: affective script, shot-cut rate, motion intensity, brightness, lighting key, color energy.	affective partition using scripts + SVM	anger, sadness, fear, joy	300 minutes	5	R: 0.836, P: 0.844
Xu et al. [48]	AR: shot duration, average motion intensity, short-time energy, MFCC; VA: brightness, lighting key, color energy, pitch.	fuzzy c-mean clustering + CRF	three level emotion intensity, fear, anger, happiness, sadness and neutral	1440 minutes	10	Accuracy: 0.807
Yoo and Cho [21]	V: average color histogram, brightness, edge histogram, shot duration, and gradual change rate	IGA	action, excitement, suspense, quietness, relaxation, and happiness	300	10	Accuracy: 0.770
Yazdani et al. [49]	A: ZCR, MFCC; V: shot boundary, lighting key, color, motion, ΔMFCC	GMM	arousal, valence, and dominance	120	32	Accuracy, arousal: 0.900, valence: 0.750, dominance: 0.720
Yazdani et al. [50]	A: ZCR, energy, MFCC; V: lighting key, shot boundary, color, motion. ΔMFCC, autocorrelation MFCC, LPC, ΔLPC, silence ratio, pitch, centroid, band energy ratio, delta spectrum magnitude	kNN	arousal and valence;	120	32 (16F,16M)	Accuracy: 0.520

A: Audio features; V: Visual features; AR: Arousal features; VA: Valence features; P: Precision; R: Recall; F: Female; M: male.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

music, and environmental sound to characterize different acoustic aspects of the video's emotion. Speech features remain the most dominant acoustic features, followed by music and environmental sounds. For speech features, continuous prosodic features like speech energy, pitch, and fundamental frequency as well as spectral features like MFCC are most commonly used for characterizing video affective content. Widely used music features include dynamics, timbre, harmony, and rhythm. In addition, many audio features are shared among different audio types including energy related features [19], [27], [38], [44], [24], [46], [2], [47], [10], [11], [15], [22], [23], [50], [31], ZCR [19], [27], [38], [11], [15], [49], [50], [31], MFCC [19], [27], [24], [11], [15], [23], [49], [50], spectral centroid [27], [24], [50], spectral rolloff [19], [27], [30], [11], spectral flux [19], [27], [30], [15], etc. These features are used for both emotional dimension prediction and emotional category classification.

3.1.2 Visual features

Early video affective content analysis focused on movies' emotion classification, which drew heavily upon methodologies from cinematography. Accepted cinematic rules and techniques, known as film grammar, are used by directors to communicate emotional cues to the audience [67], [68]. Well-established film techniques and grammar can be used to change the visual and sound elements of the movie in order to invoke or heighten the audience's emotional experience. The visual elements which filmmakers typically manipulate to inject emotion include tempo, lighting, and color.

Tempo is an important feature of films and has significant power to attract viewers' attention and to affect viewers' emotion intensity [69]. It captures the amount of camera and subject movement in each shot and between shots. According to film theorists, motion is highly expressive able to evoke strong emotional responses in viewers [70], [2]. In fact, studies by Detenber et al. [70] and Simmons et al. [71] concluded that an increase of motion intensity on the screen causes an increase in the audience's arousal. The tempo of a movie can also be changed by varying shot properties such as length and transition rate [72]. Shorter shots create a high tempo of action development and can induce stress and excitement on the part of the audiences, while longer durations create a more relaxed and slow-paced scene [73]. Furthermore, rapid shot changes can better convey dynamic and breathtaking excitement than a long and slow shot change [15]. Film tempo can also be changed by varying the camera position and movement speed in order to inject different types of emotion into the movie. For example, when the camera moves from a high shot to a low shot, the character looks imposing, which gives the feeling of authority and may create fear on the part of the audience. Camera shaking with a handheld camera can create feelings of uneasiness, danger, and stressful anticipation. A quick pushing of the camera towards the character can induce surprise and shock, while a smooth camera motion on the dolly away from the character can make the character appear lost or abandoned. A zolly camera shot (i.e. a cinematic technique that moves the camera forward or backward while changing camera zoom in the opposite direction) creates an overwhelming sense of foreshadowing [74], [75].

Lighting, the spectral composition of the light, is another powerful cinematography tool to manipulate visual elements. Lighting measures the contrast between dark and light, and influences the appearance of every element in the scene. Lighting is often exploited by directors to give a connotative signature to specifically affect the emotions of the viewer and to establish the mood of a scene [15]. Two major aesthetic lighting techniques are frequently employed: high-key lighting and low-key lighting. The former is often used to generate the lighthearted and warm atmosphere, typical of joyous scenes. In contrast, the latter uses dim lights, shadow play, and predominantly dark backgrounds to create sad, surprising, frightening, or suspenseful scenes [76], [15]. Horror movies often use low light levels, while comedies are often well lit. In addition to brightness, light and shade are used together in movies scenes to create affective effects.

Color is also an important film element that can be changed to affect the viewers' emotion. Specifically, color brightness is often used to affect valence while color saturation is used to influence arousal. Movie directors can jointly vary the valence and arousal qualities of a scene by changing the color energy. For instance, a joyous effect can be manufactured by setting up a scene with high color energy. Sad or frightening videos commonly consist of gray frames.

In summary, we can gain insight into film production rules through the cinematographic theories and principles, and use this to formulate new visual features to capture the video's affective content. Specifically, film grammar bridges the gap between the high-level semantics of movie content and low-level audio and visual features. Grammar and production rules can indicate the intended purpose or emotions that the director expects to invoke. An understanding of the specific techniques that the movie director uses allows us to design visual features accordingly. From these features we can then reverse the director's intent and hence the movie's affective content.

Inspired by film grammar, tempo, lighting and color features are extracted to characterize videos' affective content. Various features have been proposed to capture a video's tempo. Shot is an important film element that can control a video's tempo. Shot-related features [14], [11], [22], [48], [21], [49], [50] include shot duration, shot transition, and shot type transition. Shot duration represents the length of a shot. Shot transitions include cuts, fades, dissolves, and wipes. Shot type transition rate [27], [38], [44], [2], [8], [10], [15], [12], [21], [31] measures the pace variation of the employed shot type. Motion is another important film element that can control a video's tempo. Motion-related features includes motion intensity, motion dynamics, and visual excitement. Motion intensity [30], [38], [14], [22], [12], [48], [39], [31], [42] reflects the smoothness of transitions between frames. It can be estimated from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

the intensity difference of two frames [31]. Zhang et al. [31] propose to use motion intensity and shot change rate to characterize arousal. Motion dynamics [27], [44], [46], [2], [8], [10], [11], [15], [50] depends on shot pace, shot type, camera, and object motion. 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 [77]. Furthermore, image moving history [78] and motion vectors encoded in the macroblocks of MPEG [22] are proposed as motion features. Finally, features related to camera distance have also been proposed. Wang and Cheong [15] used average gray level co-occurrence matrix of a scene as a visual cue to characterize emotional distance of the scene.

Lighting-related features include light keys, tonality, etc. The lighting keys [27], [38], [24], [11], [15], [48], [49], [31] are related to two major aesthetic lighting techniques, i.e., chiaroscuro and flat lighting. The former is characterized by strong contrast between light and shadowed areas, and the later reduces the emphasis of the contrast between light and dark. Thus, for each frame the first descriptor measures the median of the pixels' brightness, while the second uses the proportion of pixels with a lightness below a shallow threshold [15]. Zhang et al. and Xu et al. [31], [48] proposed to use lighting keys to characterize arousal. Watanapa et al. [22] proposed to use tonality (the proportion of brightness to darkness in a scene) to capture the lighting in the film. Teixeira et al. [11] proposed to use the overall level of light and the proportion of shadows to capture lighting.

To better represent the movie's color, color features are typically computed in the HSV (Hue, Saturation, and Value) space, since psychological studies have shown that humans can better perceive emotions in HSV space than others [22], [79]. Typical color features include color saturation, dominant color, color layout, color energy, color heat, color activity, and color weight [18], [19], [27], [43], [24], [14], [8], [9], [8], [10], [11], [21], [49], [50]. Color saturation refers to the intensity of color. A dominant color is the most obvious color in an image, while color layout describes the spatial distribution of color in an image. Color energy [27], [38], [44], [14], [15], [22], [12], [48], [21], [31], [42] captures the perceptual strength of the color, and is affected by saturation, brightness, and location of different colors in an image. It is defined as the product of the raw energy and color contrast. Color heat is defined by the factor of warm/cool. Color activity is determined by the colour difference between a test colour and a medium gray. Colour weight is related to three factors, i. e. hard/soft, masculine/feminine, and heavy/light. Canini et al. [27] proposed to extract dominant color, saturation, color energy, color layout, and scalable color from MPEG7 video encoding. Zhang et al. [31] characterized valence using color saturation and color energy. Teixeira et al. [11] adopted color heat, color activity, and color weight to characterize a video's color.

Printed scripts can be an important clue to analyze video affective content, since scripts provide direct access to the video content. However, only recently, scripts have been introduced into emotional analysis. Xu et al. [12] proposed a two-step affective content retrieval method using scripts. First, video segments with continuous scripts are grouped as one partition, whose emotional labels are determined by emotional words in this script partition. Second, Support Vector Machine classifiers are applied to video features for affective partition validation.

3.2 Direct mapping between video content and emotional descriptors

Video features may be mapped to emotional descriptors using a classifier for categorical descriptors or a regressor for dimensional descriptors.

3.2.1 Classifications

Many machine learning methods have been investigated to model the mapping between video features and discrete emotional descriptors, including support vector machines (SVMs) [80], multi-layer feed-forward neural networks (NNs) [22], Adaboost [8], Gaussian Mixture Models (GMMs) [49], K-Nearest Neighbor (KNN) [50], Hidden Markov Models (HMMs) [10], [61], Dynamic Bayesian Networks (DBNs) [81], and Conditional Random Fields (CRFs) [48].

A classifier is divided into static or dynamic based on temporal information. Multi-layer feed-forward neural networks [22], SVMs [15], [23], [12] and GMMs [49] are used for static modeling. NNs are known to be effective for nonlinear mappings, and achieve good performance given effective features. For example, Watanapa et al. [22] proposed to classify movie clips into excitement, joy, or sadness using a two stage sieving artificial neural network, in which the first stage specialized in filtering the excitement class and the second stage classified joy and sadness. One problem with NN is its blackbox nature. Users typically do not understand its internal working mechanism. Another commonly used classifier is SVM. Because of its simplicity, its max-margin training method, and its use of kernels, SVM has achieved great success in many fields including video affective content analysis. One problem with SVM is that its selection of kernel function remains heuristic and ad hoc. Wang et al. [15] adopted a specially adapted variant of SVM to classify films into anger, sadness, fear, joy, surprise, and neutral.

Both SVM and NN are deterministic approaches. GMM is a probabilistic approach based on a convex combination of multivariate normal densities. GMMs explicitly model multi-model distributions, and are effective for emotion classification since they can capture the joint density function of multiple emotion categories in the feature space, with one Gaussian mode for each emotion category. Yazdani et al. [49] proposed to use GMM for affective content analysis of music video clips. During training, the expectation maximization algorithm is utilized, and leave-one-video-out cross-validation is adopted to find the number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

of Gaussian mixtures. Finally, eight Gaussian mixtures are created to describe each class. There is still no good way for GMM to determine the optimum number of Gaussian components. It is typically solved through trial and error or cross-validation.

SVM, NN, and GMM use the input features to perform classification only. They do not perform any feature selection. In contrast, Adaboost performs feature selection, constructs a weak classifier with each selected feature, and combines the weak classifiers to perform the final classification. As a way of dimensionality reduction, feature selection can improve not only classification accuracy but also efficiency. This is important for video affective content analysis since the input feature space may be large. Kang [8] extracted several visual features and then used an AdaBoosting algorithm to select highly meaningful features to classify emotional events. Other feature selection methods have been proposed in addition to Adaboost. Canini et al. [27] employed a filtering-based approach to select the most relevant features in terms of their mutual information to the video's affective content. Teixeira et al. [11] performed a sensitivity study to evaluate how much each basic feature contributes to the final emotion detection by removing the features one at a time and observing the performance of the model.

K-Nearest Neighbor is an effective and simple non-parametric method for classification. It has been successfully applied to many pattern recognition problems. Yazdani et al. [50] used KNN to classify music video clips into high, low, or neutral arousal; or positive, neutral, or negative valence.

Traditional single-instance learning methods such as SVM and NN require the researcher to label each training instance. However, for some videos, the emotional labels may be ambiguous or hard to ascertain. Multi-Instance Learning (MIL) [82] was introduced to handle these cases. Instead of labeling each instance as positive or negative, a set of instances, called a bag, are collectively labeled as positive or negative, with the assumption that all instances in a negative bag are negative samples and at least one instance in a positive bag is positive. A classifier is trained to classify positive and negative samples from a collection of labeled bags. MIL can hence perform learning without explicitly labeling each positive instance. Wang et al. [83] proposed to recognize horror video scenes using MIL, where the video scene is viewed as a bag and each shot is treated as an instance of the corresponding bag. Experimental results on their constructed dataset, including 100 horror and 100 non-horror video scenes collected from the Internet, demonstrate the superiority of multi-instance learning over single-instance learning for horror video scene recognition. Conventional MIL assumes the instances in a bag are independent. Therefore, it ignores contextual cues. To solve this, Ding et al. [24] further proposed a multi-view multi-instance learning model for horror scene recognition by using a joint sparse coding technique that simultaneously takes into account the bag of instances from the independent view as well as the contextual view.

Some works also employ unsupervised cluster techniques for affective video analysis, since affective states vary from person to person and it is difficult to pre-define the number and types of affective categories for a video. Xu et al. [48] adopted fuzzy c-mean clustering on arousal features to identify three levels of emotional intensity by checking the distances between sample points and each cluster center. Zhang et al. [42] used affinity propagation to cluster Music Television (MTV) videos with similar affective states into categories based on valence features and arousal features.

While the static models try to capture the mapping between input features and video emotions, they cannot effectively capture the dynamic aspects of emotion. Modeling emotion dynamics is important since emotion evolves over time. Hidden Markov Model (HMM) is the most widely used dynamic graphical model, partially due to its simple structure and its efficient learning and inference methods. Sun and Yu [10] proposed a recognition framework based on Video Affective Tree (VAT) and HMM. Four 2-state HMMs were used to model and classify joy, anger, and sadness. Kang [8] proposed to use HMM to map low-level audio-visual features to high-level emotional events by capturing the temporal patterns in the video data for each emotional state. Two HMMs of different topologies were created to classify emotions into four states: fear, sadness, joy, and normal. Xu et al. [84] developed a four-state HMM to classify audio emotional events such as laughing or horror sounds in comedy and horror videos.

The simple and restrictive topology of HMM, however, limits its representation power. In addition, determining the number of hidden states is heuristic. The Dynamic Bayesian network (DBN) is a generalization of HMMs that is also used to model video dynamics. Arifin and Cheung [18], [19] constructed an n-level DBN for affective classification in a three-dimensional space, i.e., Pleasure-Arousal-Dominance. Dynamic Bayesian networks are used to regress the emotion in the 3D emotional space from the video features. The emotional dimensions are then translated into emotion categories. Hybrid graphical models have also been proposed. Teixeira et al. [11] proposed to detect pleasure, arousal, and dominance coefficients as well as six emotion categories using two Bayesian network topologies, a hidden Markov model and an autoregressive hidden Markov model. Their system first extracts a set of low-level audiovisual features from video shots, and then feeds them into two Bayesian networks to estimate the values of pleasure, arousal, and dominance for each video segment. After that, a set of models are then used to translate the resulting Pleasure-Arousal-Dominance values into emotion categories.

Despite its powerful representation capability, DBN's complex learning and inference methods may limit its practical utility. One common problem with the existing dynamic models including HMM and DBN is that they only model local dynamics due to the underlying Markov assumption. The overall or global dynamic pattern of a time series of visual and audio signal is important in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

distinguishing between different emotions. Recent works in computer vision have begun to investigate this issue, including modeling global and high order dynamics of facial video for action unit recognition [85] and the global dynamics of body joint angle trajectories for human action recognition [86].

3.2.2 Regression

If the emotional descriptor is continuous, a regressor is used to map the features to the continuous emotional dimensions. One approach is to manually define the mapping function between low-level features and dimensional emotional descriptors. For example, Hanjalic and Xu [2] directly mapped the motion intensity, cut density, and sound energy onto the arousal dimension, and the pitch-average to the pleasure dimension by defining an analytic time-dependent function, which uses video frames for the time dimension. Chan and Jones [45] used pitch to measure the magnitude and sign of valence, and audio energy to model arousal.

Other than manually defining the mapping functions between the features and emotional dimensions, another approach is to use general regression, such as polynomial regression [38], neural network [22], or support vector regression [27], [43], [30], [38], [39], [31], [42] to learn the mapping functions from data. For example, Zhang et al. [39], [31] adopted support vector regression to map motion intensity, short switch rate, zero-crossing rate, tempo, and beat strength to arousal dimension, as well as lighting, saturation, color energy, rhythm regularity, and pitch to valence dimension. Compared with Zhang et al.'s work, Cui et al. [30], [38] extracted another three audio features from music videos, i.e., short time energy, bandwidth, and rhythm strength, and they further employed multiple linear regression and support vector regression with different kernels including exponential Radial Basis Function (RBF), Gaussian RBF, and linear and polynomial kernels for valence and arousal estimation. Instead of using valence and arousal, Canini et al. [27], [43] proposed to use natural, temporal, and energetic dimensions. They first extracted twelve visual features, sixteen audio features and three grammar features. Then, they employed an information theory-based filter, the minimum-Redundancy Maximum-Relevance scheme (mRMR), to select the most relevant features. After that, they compared three regressive procedures, i.e., polynomial combination, feed-forward neural network trained by a back-propagation algorithm, and support vector regression, for predicting the three dimensions.

3.3 Middle-level Representation

Most of the existing methods of video affective analysis directly map low-level video features to emotions by applying a machine learning method. However, there is a semantic gap between low-level features and high-level human perception of emotions, since features describe only low-level visual and audio characteristics of videos, while emotion is a high-level perception concept. In order to bridge the gap between the low-level features and the high-level emotions, recent work has introduced a middle-level representation between the video features and the video's affective content. These methods construct mid-level representations based on low-level video features and employ these mid-level representations for affective content analysis of videos.

Middle-level representation can be defined manually or learned automatically from data. Xu et al. [23] proposed a three-level affective content analysis framework by introducing a mid-level representation to capture primitive audio events such as horror sounds, laughter, and textual concepts (e.g., informative keywords). The middle-level representation is constructed from the low-level visual and audio features. The video's affective content is then inferred from the middle-level events and concepts instead of directly from the video features. Similarly, Canini et al. [27] exploited film connotative properties as the middle-level representations. Based on connotative properties and their relationships to users' personal affective preferences, users' subjective emotions are then predicted. Xu et al. [48] employed a hybrid emotion representation using both dimensional and categorical approaches. They first clustered both audio and visual features into three arousal intensities (i.e., high, medium, low). Using the arousal intensities as the middle-level representations, a CRF model further classified the emotion into types: fear, anger, happiness, sadness, and neutral.

Instead of manually defining the middle-level representations, methods have also been developed to learn the middle-level representations automatically. Acar et al. [87] proposed to learn a middle-level representation using Convolutional Neural Networks (CNNs). The middle-level representations were learned from low-level video features including MFCC and color features. The learned middle-level representations, coupled with multi-class SVMs, are then used for affective music video classification. Their experiments show that the learned mid-level audiovisual representations are more discriminative and provide more precise results than low-level audio-visual ones. Irie et al. [14] proposed to automatically capture and learn the middle-level representations using the Latent Topic Driven Model (LTDM), a variant of the Latent Dirichlet Allocation (LDA) model. Audiovisual features are extracted from the video shots and then fed into the LTDM model to automatically learn the primitive affective topics via LDA learning methods. A video's affective content is then estimated based on the affective topics as well as their distributions. In their recent work [14], Irie et al. extended their previous work by representing movie shots with Bag-of-Affective Audio-Visual Words and then applied the same LTDM architecture to generate the emotional topics as the middle-level representations, and to link the topics to emotions.

3.4 Data fusion

The two modalities in a video, i.e., visual and audio, can be fused for video affective content analysis. Data fusion can be performed in two levels: feature level and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

decision level. Feature-level fusion combines audio and video features and feeds them jointly to a classifier or regressor for video affective content analysis. Decision-level fusion, also called classifier fusion, combines the results from different classifiers. Through decision-level fusion, we can combine the merits of several classifiers while avoiding their respective limitations.

Most current works adopt feature-level fusion by concatenating audio and visual features as the input of a classifier [27], [30], [38], [24], [14], [8], [9], [8], [10], [15], [22], [12], [48], [49], [50], [39], [31], [42]. Acar et al. [87] used decision-level fusion to learn the separate middle-level representation for each modality using CNNs and fused them in the decision level. In addition to performing feature-level and decision-level fusion separately, hybrid methods have also been proposed to combine feature-level and decision-level fusion. Teixeira et al. [11] proposed to use HMM and autoregressive HMM to fuse the visual and audio features from both feature level and decision level. Their experimental results on 346 video clips demonstrated that decision-level fusion showed more balanced results compared to the feature-level fusion. Yazdani et al. [49] proposed a hybrid multilevel fusion approach, which takes advantage of both feature-level fusion and decision-level fusion. In addition to the audio and video modalities, a joint audio and video modality derived from feature fusion forms an additional modality. The final decision is granted using the sum rule over the tagging results of the three modalities. Experimental results on the music video clips used to construct the DEAP database demonstrated the superiority of the proposed hybrid fusion to both feature-level and decision-level fusion.

While feature-level fusion by simple feature concatenation is easy to implement and can be effective, it may not possible for certain features of different types and/or ranges. Feature normalization should be performed in these cases prior to concatenating features. Furthermore, feature-level fusion may create a high-dimensional feature vector that require more training data and computational cost. Decision-level fusion, on the other hand, can be more effective in these cases, since decision-level fusion deals with each type of feature independently, and only combines their classification results. In fact, decision-level fusion via classifier fusion is a well-established technique in pattern recognition and has achieved superior performance in many applications. More work should be done in this area for video affective content analysis. Finally, a hybrid fusion strategy should also be considered, since it often performs better than either feature-level or decision-level fusion alone.

3.5 Personalized video affective content analysis

Most video affective content analysis work focuses on understanding the generic content of the video, independent of the users. It is assumed that emotional reactions to a video are homogeneous across viewers. Such an assumption is unrealistic. According to the appraisal theory [88], [89], emotions are produced based on a person's subjective evaluation of a stimulus event (e.g., video) which is relevant to his/her major concerns. This suggests that people's emotions towards a video will vary, since each person's evaluation of the video stimulus may be different, depending on how important the video affective content is to his or her central concerns. Moreover, according to the appraisal theory, each person varies in his or her ability to regulate emotion. This variation in regulation may lead to differences in expressing and responding to emotions. As a result, a viewer may deny, intensify, or weaken his or her true emotion. This could cause problems with the integrity of the reported emotion labels.

Given this understanding, Hanjalic and Xu [2] showed that there are two kinds of emotional descriptors: the expected emotion and the actual emotion. The expected emotion is the emotion that the video makers intended to communicate to a viewer. It can be thought of as a video's generic emotion label. In contrast, the actual emotion is the emotional response of a particular user to the video. It is context-dependent and subject-specific, and it varies from one person to another. It is an individual emotion. The goal of personalized affective content analysis is to estimate a video's personal affective label.

Recent research is moving towards user-specific/personalized video affective content analysis. In fact, this kind of personalized affective analysis is gaining increasing attention in many user-oriented applications, since personalized affective analysis can achieve better performance and improve usability. Zhang et al. [31] introduced an integrated system for personalized music video affective analysis by incorporating a user's feedback, profile, and affective preference. Canini et al. [27] exploited film connotative properties as middle-level representations. A film's connotative properties capture conventions that help invoke certain emotions in the viewers. The user's subjective emotional response is predicted based on connotative properties and their relationships to the user's personal affective preferences. Yoo and Cho [21] adopted an interactive genetic algorithm to realize individual scene video retrieval. Wang et al. [90] employed a Bayesian network to capture the relationships between the video's common affective tag and the user's specific emotion label, and used the captured relationship to predict the video's personal tag. Soleymani et al. [91] tackled the personal affective representation of movie scenes using the video's audiovisual features as well as physiological responses of participants to estimate the arousal and valence degree of scenes. Finally, implicit video tagging, which we will discuss in the next section, infers a video's affective content based on the user's spontaneous non-verbal responses. Since the response to the same video can vary between users, implicit video tagging is more personal and subjective.

4 Implicit Video Affective Content Analysis

Unlike the direct approach, in which the emotional tagging of a video is inferred from the visual and audio features,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

implicit video affective content analysis performs automatic analysis of users' reactions to infer the video's affective content. Pantic and Vinciarelli [92] are the first to introduce the concept of implicit human-centered tagging.

Currently, implicit video affective content analysis mainly adopts users' physiological signals and spontaneous visual behaviors, since most emotion theories [93][94] agree that physiological activity is an important component of emotional experience, and facial expression is the primary channel for emotion communication. Physiological signals reflect unconscious body changes, and are controlled by the sympathetic nervous system (SNS), while facial behaviors can be adopted voluntarily or involuntarily. Thus, physiological signals provide more reliable information for emotions compared to facial behaviors. However, to obtain physiological signals, users are required to wear complex apparatuses, while only one remote visible camera is needed to record facial behaviors. Furthermore, physiological signals are sensitive to many artifacts, such as involuntary eye-movements and irregular muscle movements. These body conditions rarely disturb users' facial behaviors. Such freedom of motion makes users feel comfortable to express their emotions. Therefore, compared with physiological signals, spontaneous visual behavior is more convenient and unobtrusive to measure, although it is susceptible to environmental noise, such as lighting conditions, occlusion, etc..

Recently, Soleymani and Pantic [5] have reviewed human-centered implicit tagging on images, videos, and search results. In the sections below, we review current research of implicit video content analysis. Unlike their survey, which lists all work together [5], we categorize current research into three groups based on adopted modalities: physiological signals, spontaneous visual behavior, and both. Table 3, Table 4, and Table 5 list the current research of implicit video affective content analysis from physiological signals, visible behavior, and both modalities respectively, together with adopted modalities and features, classifiers, emotion descriptors, the size of used datasets **(i.e., the number of video clips if not explicitly stated)**, the number of subjects, and experimental results.

4.1 Implicit video affective content analysis using physiological signals

Some researchers have focused on implicit video affective content analysis using physiological signals, including electroencephalography (EEG), electrocardiography (ECG), electromyography (EMG), skin temperature (ST), Galvanic Skin Response (GSR or Electrodermal Response: EDR; or Electrodermal Activity: EDA), Heart Rate (HR), and Blood Volume Pulse (BVP), as shown in Table 3.

Money and Agius [26] are among the first to investigate whether users' physiological responses can serve as summaries of affective video content. Five physiological response measures, i.e. GSR, respiration, BVP, HR and ST, were collected from 10 subjects during their watching of three films and two award-winning TV shows. By assessing the significance of users' responses to specific video subsegments, they demonstrated the potential of the physiological signals for affective video content summaries. Based on their study, they [101] further proposed Entertainment-Led Video Summaries (ELVIS) to identify the most entertaining video sub-segments. Soleymani et al. [91] analyzed the relations between video content, users' physiological responses, and emotional descriptors with valence and arousal using correlation analysis and linear relevance vector machine. Experimental results on a dataset of 64 scenes from eight movies watched by eight participants demonstrated that in addition to video features, subjects' physiological responses (i.e., GSR, EMG, blood pressure, respiration, and ST) could provide affective ranking to video scenes. Fleureau et al. [97] proposed a two-stage affect detector for video viewing and entertainment applications. They first recognized affective events in the videos, and then use Gaussian processes to classify the video segments as positive or negative using GSR, heart rate, and electromyogram. Three realistic scenarios, including mono-user, multi-user, and extended multi-user simulations, were conducted to evaluated the effectiveness of the detector on a dataset of 15 video clips viewed by 10 users. Chêne et al. [96] proposed a user-independent video highlight detection method using inter-users' physiological linkage, which is calculated from EMG, BVP, EDA, and skin temperature. Experiments on a dataset 26 scenes viewed by eight users demonstrated the validity of the proposed system.

While these researchers verified the feasibility of several physiological signals as implicit feedback, other researchers focused on only one or two physiological signals. For example, Canini et al. [32] investigated the relationship between GSR and arousal of videos using correlation analysis. Their experiments on a dataset of eight subjects watching four video clips demonstrate a certain dynamic correlation between arousal indicated by GSR and video content. Smeaton and Rothwell [104] tried to detect film highlights from viewers' HR and GSR by comparing the physiological peaks and the emotional tags of films. Their experiments on a database of six films viewed by 16 participants showed high correlation between subjects' physiological peaks and emotional tags of videos, and the catalysis of music-rich segments in stimulating viewer response. Toyosawa and Kawai [105] proposed heart rates for video digesting. They determined the attention level of each segment through deceleration of heart rate and the high frequency component of heart rate variability. Their experiments on a dataset collected from 10 subjects during watching three videos demonstrated the effectiveness of the proposed two attention measures for arousing and event-driven contents, but not for story centric videos and monotonous videos with weak arousal and neutral valence. Chen et al. [13] proposed XV-Pod, an emotion aware affective mobile video player, to select videos based on the recognized users' emotional state from GSR and heart flux. Their empirical study on four participants watching funny and upsetting video clips demonstrated that GSR and Heat Flux are two

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

TABLE 3: Implicit video affective content analysis from physiological signals

References	Modalities and features	Analysis methods	Emotion descriptors	♯**Video clips**	♯Subjects	Results
Abadi et al. [95]	MEG	Naive Bayesian Classifier	arousal and valence	32+40	7 (3F,4M)	Accuracy: music videos: 0.657; movies: 0.656
Canini et al. [32]	GSR	analyze the relations between arousal and feature using correlation coefficient	continuous arousal	4	8	GSR correlates with affective video features. Subjective scores correlates with GSR
Chêne et al. [96]	EMG, BVP, EDA, ST	sliding-window correlation + SVM	highlights	26	8	Accuracy: 0.782
Chen et al. [13]	GSR, and heat flux	Decision Tree	happiness, sadness and boredom	8	4	Accuracy: 0.928
Fleureau et al. [97]	GSR, facial EMG, PPG	Gaussian Process Classifiers	valence	15	10 (2F,8M)	Specificity: 0.859; Sensitivity: 0.872; Accuracy: 0.873
Kierkels et al. [34]	ECG, GSR, respiration, ST	2D Gaussian probability distribution.	arousal	64	7	P: 0.400
Koelstra et al. [98]	N400 ERP	ANOVA	correct or incorrect tag	49	17 (5F,12M)	congruent and incongruent tags can be distinguished by N400 activation.
Koelstra et al. [99]	GSR, BVP, respiration, ST, EMG and EOG	SVM	arousal and valence	70	6	Accuracy: 0.855
Krzywicki et al. [100]	middle-wave infrared image	analyze the relations between temperature pattern changes and visual and auditory stimuli	amusement, anger, guilt, happiness, interest, joy, embarrassment, fear, sadness,shame, surprise, unhappiness, love, pride, confusion, contempt, and disgust	3	10 (4F,6M)	there exist explicit thermal responses from the face due to physiological emotion
Money et al. [25], [26]	EDR, HR, BVP, RR, respiration rate and amplitude	Define percentile rank user response values	action/sci-fi, horror/thriller, comedy, drama/action, and drama/comedy	5	10 (4F,6M)	percentile rank user response values:80%
Money et al. [101]	EDR, HR, BVP, RR, respiration rate and amplitude;	proposed ELVIS technique,	comedy, horror/comedy, and horror	3	60	percentage overlap scores: 0.460
Soleymani et al. [102], [91], [35]	GSR, BVP, RSP, ST, EMG zygomaticus and frontalis, eye blinking rate	linear relevance vector machine	continuous valence and arousal	64	8 (3F,5M)	Mean squared error: VA: 0.014, AR: 0.027.
Soleymani et al. [103]	EEG	SVM and Linear Discriminant Analysis	three-level arousal and three level valence ; relevant or non-relevant	MAHNOB-HCI	50	F1: 0.890 for arousal, 0.830 for valence; Accuracy: 0.830
Smeaton et al. [104]	HR and GSR	compare the physiological peaks with the emotional descriptors	Salway's 21 emotion types	6	16	people experience similar and measurable physiological responses to emotional stimuli in films.
Toyosawa et al. [105]	HR	define two attention measures from heart activity	valence and arousal	3	10	P: 0.430
Yazdani et al. [17]	EEG	Bayesian Linear Discriminant Analysis	joy, sadness, surprise, disgust, fear, and anger	24+6	8	Accuracy: 0.802

good indicators of users' emotional response with arousal and valence. They further proposed to recognize emotions from GSR and heart flux using decision tree. Yazdani et al. [17] proposed to perform implicit emotion multi-media tagging through a brain-computer interface system based on a P300 event-related potential (ERP). The experimental results on a dataset of 24 video clips and six facial expression images watched by eight subjects showed the effectiveness of their system. Furthermore, naive subjects who have not participated in training phase can also use the system efficiently.

Instead of using contact and intrusive physiological signals, Krzywicki et al. [100] analyzed facial thermal signatures, a non-intrusive physiological signal, in response to emotion-eliciting film clips. By comparing the distribution of facial temperatures to the summarized video clip events from a dataset of 10 subjects viewing three film clips, they concluded that different facial regions exhibit different thermal patterns, and that global temperature changes are consistent with stimulus changes. Abadi et al. [95] introduced the magnetoencephalogram (MEG), a non-invasive recording of brain activity, to differentiate between low versus high arousal and low versus high valence using naive Bayes classifier. They collected MEG responses from seven participants watching 32 movie clips and 40 music videos. The experimental results verified the feasibility of MEG signal for encoding viewers' affective responses.

Other than analyzing affective content of videos, Koelstra et al. [98] attempted to validate video tags using an N400 ERP. The experimental results on a dataset with 17 subjects, each recording for 98 trials, showed a significant difference in N400 activation between matching and non-matching tag.

Although the studies described above explored physiological signals to analyze the affective content of a video, their purposes (i.e., summarization [26], [105], retrieval [91], highlight detection [96], and tagging [104], [106]) are not the same. In addition, they used different methods for emotion dimension prediction, such as linear relevance vector machine [91], and emotion category classification, such as decision tree [13], Gaussian process classifiers [97] and SVM [96], [106] etc.

These studies illustrate the development of methods for using physiological signals in the implicit video affective content analysis. However, to acquire physiological signals, subjects are usually required to wear several contact apparatuses, which may make them feel uncomfortable and hinder the real application of these methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

Recently, great progress has been made in wearable devices and remote sensing, such as Motorola Moto 360, Microsoft Kinect, and Google Glass. These advanced sensors may reduce the intrusiveness of traditional physiological sensors and cameras. We believe that video affective content analysis will attract more attention as these built-in sensors become available.

4.2 Implicit video affective content Analysis using spontaneous visual behavior

Some researchers have focused on implicit video affective content analysis from human spontaneous facial behavior, as shown in Table 3.

Joho et al. [107] proposed to use viewers' facial activities for highlight detection. The experimental results on a dataset of six participants watching eight video clips suggested that compared with the activity in the lower part of the face, the activity in the upper part of the face is more indicative of personal highlights. Zhao et al. [20] extracted viewers' facial expressions frame by frame, and then obtained the affective curve to describe the process of affect changes. Through the curve, they segmented each video into affective sections.

Rather than focusing on subjects' whole facial activity, Ong and Kameyama [16] analyzed affective video content using viewers' pupil sizes and gazing points. Katti et al. [33] employed users' pupillary dilation response for affective video summarization and storyboard generation.

Peng et al. [109] proposed to fuse the measurements of users' eye movements and facial expressions for home video summarization. Their experimental results on eight subjects watching five video clips verified that both eye movements and facial expressions can be helpful in video summarization.

In addition, two studies demonstrate the feasibility of using facial responses for content effectiveness evaluation. McDuff et al. [108] proposed to classify liking and desire to watch again automatically from spontaneous smile responses. Yeasin et al. [111] employed the recognized facial expressions to detect interest levels.

Most studies of implicit video affective content analysis assume that the expressions displayed by the subjects accurately reflect their internal feelings when they watched the videos. Therefore, most researchers have directly used the recognized expression as the emotional descriptor of the videos. However, users' internal feelings and displayed facial behaviors are not always the same, evern though they are related [112], since expressions of emotion vary with person and context. Wang et al. [90] employed a Bayesian network to capture the relationship between facial expressions and inner personalized emotions, and thus improved the performance of video emotion tagging.

All these studies have demonstrated the potential for using spontaneous visual behavior for implicit video affective content analysis. Their purposes cover summarization [33], [107], [109], [110], [113], recommendation [109], [114], and tagging [16], [108], [111], and the used modalities include facial expression [107], [109], [114], [109], [110], [108], [111], click-through action [109], [114], and eye movement [16], [33], [109], [110], [109]. In addition, they used different methods for facial expression recognition, including eMotion (a facial expression recognition software) [109], [114], SVM [109], [110], Bayesian networks [107], [115], HMM [111] and hidden conditional random field [20].

4.3 Hybrid methods

In addition to using physiological signals or visual behavior signals independently, there is also work that combines the two modalities. Table 5 summarizes the related hybrid methods. Koelstra et al. [28] proposed to use both facial expressions and EEG signals for video affective content analysis in the valence-arousal space. Experimental results on the MAHNOB dataset [116] demonstrated that both feature-level and decision-level fusion methods improve tagging performance compared to single modality, suggesting the modalities contain complementary information. The same conclusion is also reached by Soleymani et al. [117], [29], who proposed to recognize emotions in response to videos by fusing users' EEG signals, gaze distance, and pupillary response from both feature-level fusion and decision-level fusion. As well as analyzing affective content of videos, Arapakis et al. [118] verified the feasibility of topic relevance prediction between query and retrieved results from audiences' facial expression and peripheral physiological signals such as GSR and skin temperature.

5 Benchmark Database

Most video affective content analysis research requires emotion-induced video databases. Constructing a benchmark video database for emotion tagging and analysis is a key requirement in this field. However, as seen in Table 1 to Table 5, we find that most current works constructed their own databases to validate their methods.

For direct video affective content analysis, Baveye et al. [119] proposed the LIRIS-ACCEDE, which is composed of 9800 video clips extracted from 160 movies shared under Creative Commons licenses. Therefore, this database is publicly available without copyright issues. 1,517 annotators from 89 different countries participated in the rating-by-comparison experiments on crowdsourcing. FilmStim [120] is another database including 70 film excerpts annotated by 364 participants using 24 emotional classification criteria. The MediaEval 2010 Affect Task corpus [121] was constructed for boredom detection. It includes 126 videos between two and five minutes in length, which are selected from a travelogue series called My Name is Bill, created by filmmaker Bill Bowles. The videos, the extracted speech, and the available metadata including the episodes' popularity and the ground truth annotations obtained from the crowdsourcing platform Mechanical Turk are provided. Recently, Tarvainen et al. [122] constructed a dataset consisting of 14 movie clips 1-2.5 minutes in length, rated by 73 viewers with 13 stylistic attributes, 14 aesthetic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

TABLE 4: Implicit video affective content analysis from visible behavior

References	Modalities and features	Analysis methods	Emotion descriptors	♯**Video clips**	♯Subjects	Results
Joho et al. [107]	facial expressions	Naive Bayesian Classifiers	highlights	8	6 (3F, 3M)	F-score: 0.350
Katti et al. [33]	pupil and eye gaze	define the relation between the arousal and pupillary dilation	continuous arousal	5	20	Verify the effectiveness of pupillary dilation for affective video content analysis
McDuff et al. [108]	feature points and local binary patter features	Naive Bayesian, SVM, HMM, and Hidden-state and Latent Dynamic CRF	liking and desire to watch again	3	6729	AUC: 0.800, 0.780
Ong et al. [16]	pupil features and gazing points features	K-means cluster	neutral, happiness and horror	133	6 (3F, 3M)	Accuracy: 0.719
Peng et al. [109], [110]	head motion, center of an eyeball, two corners of the eye and the upper eye lid and facial expression	define the relations between interest and features	interested or not interested	5	8 (2F, 6M)	The proposed Interest Meter can recognize shot clips that interest subjects.
Wang et al. [90]	head motion and facial expression	SVM + BN	happiness, disgust, fear, surprise, anger, sadness	NVIE	about 100	Accuracy: 0.548
Yeasin et al. [111]	facial expression	HMM + k-NN	arousal, valence and stance	488	97	Accuracy: 0.909
Zhao et al. [20]	facial expression	hidden CRF	comedy, tragedy, horror, moving, boring, exciting	100	10 (4F, 6M)	Accuracy: 0.854

TABLE 5: Implicit video affective content analysis from both physiological signals and visible behavior

References	Modalities and features	Analysis methods	Emotion descriptors	♯**Video clips**	♯Subjects	Results
Arapakis et al. [118], [109], [114]	facial expressions, ST, GSR, click-through, HR, acceleration, heat flux	SVM; KNN	relevant and irrelevant	200 hour videos	24 (10F,14M)	Accuracy: 0.665; P: 0.662; R: 0.676
Koelstra et al. [28]	EEG, facial expression	Gaussian Naive Bayesian classifier	discrete valence, arousal and control	MAHNOB-HCI	30	Accuracy: arousal: 0.800, valence: 0.800, control: 0.867
Soleymani et al. [117], [29]	pupil diameter, gaze distance, eye blinking, EEG.	feature-level: SVM; decision-level: probabilistic model	three level arousal and three level valence	20	24	Accuracy: feature-level:(0.664, 0.584), decision-level:(0.764, 0.685); F1: feature-level: (0.650, 0.550), decision-level: (0.760, 0.680)

attributes, and temporal progression of both perceived and felt valence and arousal.

For implicit video affective content analysis, Koelstra et al. constructed DEAP (Database for Emotion Analysis using Physiological signals) [35]. It includes electroencephalogram and peripheral physiological signals (i.e., GSR, respiration, ST, ECG, BVP, EMG, and electrooculogram (EOG)), collected from 32 participants during 40 one-minute long excerpts of music videos. It also includes frontal face videos from 22 participants. The videos are annotated in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity. The MAHNOB-HCI [116] database consists of face videos, speech, eye gaze, and both peripheral and central nervous system physiological signals from 27 subjects during two experiments. In the first experiment, subjects self-reported their emotional responses to 20 emotion-induced videos using arousal, valence, dominance, and predictability as well as emotion categories. In the second experiment, subjects assessed agreement or disagreement of the displayed tags with the short videos or images. The NVIE (Natural Visible and Infrared Facial Expression) database [123] is another multimodal database for facial expression recognition and emotion inference. It contains both posed expressions and video-elicited spontaneous expressions of more than 100 subjects under three illumination directions. During the spontaneous expression collection experiments, the participants self-reported their emotion experience to the stimuli video according to six basic emotion categories, namely happiness, disgust, fear, sadness, surprise, and anger. None of the three databases provide the stimulus videos to the public due to copyright issues. However, researchers can directly obtain the information from the database constructors. Recently, Abadi et al. [124]proposed DECAF (a multimodal dataset for DECoding user physiological responses to AFfective multimedia content) database. The database comprises synchronously recorded MEG data, near-infrared facial videos, horizontal Electrooculogram (hEoG), ECG, and trapezius-Electromyogram (tEMG) peripheral physiological responses from 30 participants watching 40 one-minute music video segments and 36 movie clips. Participants report their assessment of valence and arousal ratings for music and movie clips.

6 Challenges and Recommendations for Future Directions

Although much work has been conducted on video affective content analysis in recent years, and great progress has been made, video affective content analysis remains in its infancy. There are many challenges in three aspects: emotion annotation for emotional descriptors, feature extraction (i.e.,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

representation) for characterizing video affective content as well as measurements of users' physiological or visual behavior response, and relations among video content, users' response and emotional descriptors. In this section, we briefly discuss the challenges and potential solutions in each of these areas.

6.1 Emotional descriptors

Almost all research of video affective analyses adopt self-reported data for the ground truth labels. However, the emotion of the subjects is very difficult to obtain, and even self-reports are not always reliable due to many problems such as cognitive bias [125]. Recently, Healey's work [126] indicated that triangulating multiple sources of ground truth information, such as in situ rating, end-of-day rating and third party rating, leads to a set of more reliable emotion labels. We may refer to this work to obtain ground truth emotion labels in future research.

Most present research has assumed that there is only one emotional descriptor or point in emotional dimensional space for a video. However, it is very hard to find videos that induce a high level of a single emotional category without the presence of other emotions, either in day-to-day living or inside the laboratory. Research [127], [128] demonstrates that when users watch amusing videos, they often feel amused, happy, and surprised simultaneously. The videos that induce anger may also induce some degree of disgust, sadness, fear, and/or surprise, but not high levels of happiness. Therefore, emotional annotation should accommodate the simultaneous presence of multiple emotions. Furthermore, the co-existent and mutually exclusive relations among emotions should be exploited during video affective content analysis [34], [129].

Finally, there is also the issue of video granularity. The smallest analyzable unit for current video affective content analysis can be a shot, a scene, or an arbitrary segmentation. Some research uses a shot as the smallest analyzable unit. The assumption is that emotions do not change during a shot. Subsequently, for each shot, there should be one characteristic emotion that the audience is expected to feel. However, a shot (i.e., an uninterrupted run of a camera take) may not be adequate to convey a concept and/or induce a well-defined emotional state in the user, since its average length is short. Instead, a scene in a movie depicts a self-contained high-level concept. That is why some works chose scene as the elementary unit. However, emotions may change across scenes [130]. Other works adopted arbitrary segmentation as the elementary unit. The length of the segments varies from several seconds to several minutes. Since the duration of users' emotional state varies with the stimulus videos and users, the granularity of video affective content analysis may also vary. An adaptive emotion change model may be helpful to select the proper granularity.

6.2 Feature representation for video content and users response

Current research uses hand-crafted features to represent the video's affective content. Most existing features are inspired by cinematic principles and concepts, while most features for characterizing users' emotional responses are inspired by psychological research. Since the relationship between the low-level video features and users' emotional responses is still not well understood, and may varies by task and by individual, it may be better to automatically learn features from data instead of using the hand-crafted features. Deep learning [131] has been quite effective in learning feature representation. It has achieved excellent performance in several fields including computer vision [132], speech recognition [133], and brain-computer interfaces [134]. Unlike shallow learning, the advantages of deep learning include hierarchical data representations of features at different levels of abstraction that match better with the human visual cortex, and are more tolerant to image variation including geometric and illumination variation, yielding greater predictive power. Therefore, future research should investigate deep learning or other feature learning methods to learn more effective features to characterize video's affective content and to characterize user's emotional responses.

Moreover, instead of performing purely data-driven feature learning, we should not ignore years of research when deriving hand-crafted video features based on cinematography and physiology. This can be accomplished by incorporating cinematographic and physiological knowledge into current data-based feature learning methods. This line of research has achieved promising results in several areas including brain-computer interaction (BCI) [135] and represents a brand new direction to pursue for video affective content analysis.

Furthermore, video affective content analysis via a middle-level representation is a promising direction to pursue in the future. This research can be applied to both direct and implicit video affective content analysis. Middle-level representation can better bridge the semantic gap between the low-level features and high-level users' emotions. Through middle-level representation, the relationships between video features and emotion descriptors as well as users' spontaneous responses and emotion descriptors can be better modeled, leading to more accurate predictions as demonstrated by the existing work in this area. In addition to further extending the existing middle-level representations, one possible new direction in this area is to combine the manually defined middle-level representation with the automatically learned middle-level representation so that semantically meaningful middle-level representations can be learned from the data.

6.3 Hybrid analysis

Present direct video affective content analysis approaches involve mapping from video content to emotional descriptors. Implicit video affective content analysis approaches map users' spontaneous responses to emotional descriptors. Little research considers the relationships among video content, users' spontaneous responses, and users' emotional states simultaneously. We believe that fully exploiting the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

three aspects and their relationships is crucial to reducing the semantic gap between the low-level video features and the users' high-level emotional descriptors. Thus, hybrid analysis from both video content and viewers' responses may improve tagging performance [3]. Furthermore, since subjects may be uncomfortable by wearing sensors to detect their body changes or being observed by cameras during the actual tagging, it is more practical to employ users' physiological responses during model training only [136]. In actual tagging, only video features would be used, without measuring user response [136]. The study of learning using information that is available only during training, i.e., the privileged information [137], will be a potential solution. This may be referred to as implicit hybrid analysis.

6.4 Personalized video affective content analysis

Since assessment of the affective content of a video varies greatly from person to person, it is increasingly necessary to perform personalized video affective content analysis. This requires the individualization of the current affective analysis models by incorporating a user's personal profile and affective preferences into the models. As mentioned above, such a personalized affective analysis can achieve better performance and improve usability. One potential solution to better model individual emotion is the multi-task learning technique. With multi-task learning, the personal emotion for one viewer can be regarded as a single task representing an individual factor, while multiple individual emotion models for multiple viewers share common factors. By learning the multiple individual emotion models simultaneously and exploiting their commonality, multi-task learning can lead to an improved individual emotion model.

6.5 Context and prior knowledge

Present research has ignored the contextual nature of emotion. However, the same viewer can experience different emotions in response to the same video depending on the context, including time, temperature, mood [138], degree of familiarity, motivation, and social context. Important contextual factors should be incorporated into the emotional analysis process. One possible approach is to use probabilistic graphical models, which can fully integrate contextual information and observable evidence for emotional analysis [139].

In addition to context, current analysis ignores some prior knowledge that is readily available, particularly during training. For example, current video affective content analysis does not consider gender differences in response to emotional stimuli. The neurophysiological study by C. Lithari et al. in [140] proved that there are gender differences in response to emotional stimuli. This observation may provide prior knowledge in video affective content analysis. For example, the prior probability of emotions for male and female subjects may be different for the same videos. Thus, gender differences could be exploited during training as privileged information [137] to produce a better classifier or regressor, or be exploited as prior information during testing to improve the performance of video affective content analysis.

6.6 Actor-centered video affective content analysis

Existing video affective content analysis focuses on generic video elements, but both generic video elements and actor-related specific video elements can affect the viewers' emotions. Less work has been done on actor properties and their effect on viewers' emotions. Actor attributes are important since viewers may empathize with and attach emotions to the actors in the videos. In fact, study of film syntax suggests the performance of actors and actresses, including actor-related attributes such as gesture or facial expression, can effect a viewer's emotion. Actor-centered video affective content analysis is, in particular, important for user-generated videos since ordinary users usually do not have the expertise to apply various film grammars to introduce viewers' emotions. Srivastava et al. [141] addressed the recognition of emotions of movie characters. Low-level visual features based on facial feature points are employed for facial expression recognition, whereas lexical analysis of dialog is performed in order to provide complementary information for the final decision. Actors in a video typically express their emotions through non-verbal behaviors such as facial expression and body gestures. Actor-centered video affective content analysis can therefore benefit from the large body of research on human expression and body gesture recognition in computer vision.

6.7 Benchmark Database

Almost all studies of video affective content analysis adopt their own individually developed corpus, typically with small size and little information on context and individual emotional descriptors. Creating a larger scale high-quality database is a prerequisite to advance algorithm development for video affective content analysis [119]. However, it is not only time-consuming but also difficult, since emotions are subjective and context-dependent, and emotional labels are usually collected from viewers' self-reports. Soleymani et al. pointed out three critical factors for affective video corpora: the context of viewer response, personal variation among viewers, and the effectiveness and efficiency of corpus creation [142]. In order to model emotional responses that vary with context and across videos, the database should include a large number of responses collected from a very large and representative population. Online systems and crowdsourcing platforms (e.g., the Mechanical Turk) may be exploited to recruit large numbers of viewers with a representative spread of backgrounds to watch videos and provide contextual information on their emotional responses. Both the common and personalized emotion tags should be collected. Finally, we briefly discuss the annotation reliability issue. Data labeling and annotation is a subjective process and the results may vary with the expertise level of the annotators.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2432791, IEEE Transactions on Affective Computing

Poor annotation could adversely affect the subsequent video affective content analysis. To improve annotation quality and consistency, multiple annotators should be employed and the final annotation results should reflect the consensus among different annotators, through triangulating multiple sources of ground truth information [126] or more sophisticated techniques such as the MACE technique introduced in [143] that uses an unsupervised item-response model to infer the underlying true annotation from different individual annotators. This issue deserves further investigation in the video affective content analysis community.

7 Conclusion

In this paper, we provide a thorough review of current research of video affective content analysis. We first lay out the major components for video affective content analysis, This is followed by a discussion of different emotion theories and their definitions of emotions as well as the relationships among emotion definitions. We review related work for two major video affective content analysis approaches **with a focus on the direct video affective content analysis.** For direct video affective content analysis, we review the audio and visual features used to characterize the video's affective content. We, in particular, focus on the video features inspired by the cinematic principles and physiological studies. We classify video features into different categories and compare their strengths in characterizing different aspects of video emotion and different types of video emotion. We then review different types of models that map video features to a video's affective contents. We group the models into static models and dynamic models, as well as probabilistic models and deterministic models, and compare their respective strengths and weaknesses. In addition to discussing well-established topics, we also discuss emerging topics on middle-level representations and personalized video affective content analysis. For implicit video affective content analysis, we provide a review of video affective tagging from physiological signals, visual behavior, and both. We also review the available benchmark databases for both direct and implicit video affective content analysis.

Finally, we identify research issues, future directions and possible solutions for several areas of video affective content analysis. For emotion data annotation, we point out the need for providing multiple emotion labels for each video due to the simultaneous presence of multiple emotions. We also recommend capturing and exploiting the relationships among multiple emotions to improve emotion recognition performance.

We also identify a few research issues for direct video affective content analysis. For video feature representation, we propose to investigate the latest deep learning paradigm to automatically learn video features to effectively characterize a video's affective content. We further propose to augment the data-driven deep learning with the established physiological and cinematic knowledge in order to avoid over-fitting and to improve the generalization ability of the learnt features. Furthermore, to better bridge the semantic gap between video features and its emotional content, we recommend further research on the middle level representation. Specifically, we propose a hybrid middle level representation that combines manually derived mid-level representation with those learnt automatically from the data. **Since the affective preferences vary with users, video affective content analysis should be personalized to fit to each person's specific affective preferences.** For personalized video affective content analysis, we propose to use multi-task learning methods to perform individualized video affective content analysis. Finally, to perform affective content analysis on user-created videos, we propose to focus on actor-specific video features since the traditional physiologically and cinematically inspired features for films and movies may not be applicable to amateur videos.

For the implicit video affective content analysis, future research should focus on using less intrusive wearable sensors to acquire physiological signals as well as on combining sensors of different modalities to measure user's spontaneous responses.

References

[1] F. Nack, C. Dorai, and S. Venkatesh, "Computational media aesthetics: finding meaning beautiful," *MultiMedia, IEEE*, vol. 8, no. 4, pp. 10–12, 2001.

[2] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Trans.*, vol. 7, no. 1, pp. 143–154, 2005.

[3] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users EEG and video content," *Multimedia Tools and Applications*, pp. 1–27, 2013.

[4] S. Wang and X. Wang, *Kansei Engineering and Soft Computing: Theory and Practice, Chapter 7: Emotional Semantic Detection from Multimedia: A Brief Overview*. Pennsylvania: IGI Global, 2010.

[5] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *IEEE Int'l Conf. Systems, Man and Cybernetics, 2012*, 2012, pp. 3304–3309.

[6] P. Ekman, *Handbook of Cognition and Emotion*. Sussex, UK: John Wiley, 1999, ch. Basic Emotions, pp. 45–60.

[7] W. M. Wundt, *Grundzüge de physiologischen Psychologie*. Leipzig: Engelman, 1905.

[8] H.-B. Kang, "Affective content detection using HMMs," in *Proc. the 11th ACM int'l Conf. Multimedia*, NY, USA, 2003, pp. 259–262.

[9] ——, "Emotional event detection using relevance feedback," in *Image Processing (ICIP 03). Int'l Conf.*, vol. 1, 2003, pp. I–721–4 vol.1.

[10] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *Affective Computing and Intelligent Interaction*, 2007, pp. 594–605.

[11] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools and Applications*, vol. 61, no. 1, pp. 21–49, 2012.

[12] M. Xu, X. He, J. S. Jin, Y. Peng, C. Xu, and W. Guo, "Using scripts for affective content retrieval," in *Advances Multimedia Information Processing*, Shanghai, China, 2010, pp. 43–51.

[13] X. Y. Chen and Z. Segall, "XV-Pod: An emotion aware, affective mobile video player," in *Computer Science and Information Eng., 2009 WRI World Congress*, 2009, pp. 277–281.

[14] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawaki, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *Trans. Multi.*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[15] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *Circuits and Systems for Video Technology, IEEE Trans.*, vol. 16, no. 6, pp. 689–704, 2006.

[16] K.-M. Ong and W. Kameyama, "Classification of video shots based on human affect," *Information and Media Technologies*, vol. 4, no. 4, pp. 903–912, 2009.

[17] A. Yazdani, J.-S. Lee, and T. Ebrahimi, "Implicit emotional tagging of multimedia using EEG signals and brain computer interface," in *Proc. 1st SIGMM Workshop Social media*, 2009, pp. 81–88.

[18] S. Arifin and P. Y. K. Cheung, "A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information," in *Proc. the 15th int'l Conf. Multimedia*, 2007, pp. 68–77.

[19] S. Arifin and P. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *Multimedia, IEEE Trans.*, vol. 10, no. 7, pp. 1325–1341, 2008.

[20] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *Proc. 19th ACM Int'l Conf. Multimedia*, 2011, pp. 1473–1476.

[21] H.-W. Yoo and S.-B. Cho, "Video scene retrieval with interactive genetic algorithm," *Multimedia Tools and Applications*, vol. 34, no. 3, pp. 317–336, 2007.

[22] S. C. Watanapa, B. Thipakorn, and N. Charoenkitkarn, "A sieving ANN for emotion-based movie clip classification," *IEICE trans. information and systems*, vol. 91, no. 5, pp. 1562–1572, 2008.

[23] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *Multimedia Tools and Applications*, pp. 1–23, 2012.

[24] X. Ding, B. Li, W. Hu, W. Xiong, and Z. Wang, "Horror video scene recognition based on multi-view multi-instance learning," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 599–610.

[25] A. G. Money and H. Agius, "Feasibility of personalized affective video summaries," in *Affect and Emotion Human-Computer Interaction*, 2008, pp. 194–208.

[26] ——, "Analysing user physiological responses for affective video summarisation," *Displays*, vol. 30, no. 2, pp. 59–70, 2009.

[27] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *Circuits and Systems for Video Technology, IEEE Trans.*, vol. 23, no. 4, pp. 636–647, April 2013.

[28] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[29] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *Affective Computing, IEEE Trans.*, vol. 3, no. 2, pp. 211–223, 2012.

[30] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. Jin, "Mutual information-based emotion recognition," in *The Era of Interactive Media*. Springer New York, 2013, pp. 471–479.

[31] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *Multimedia, IEEE Trans.*, vol. 12, no. 6, pp. 510–522, 2010.

[32] L. Canini, S. Gilroy, M. Cavazza, R. Leonardi, and S. Benini, "Users' response to affective film content: A narrative perspective," in *Content-Based Multimedia Indexing (CBMI 10), 2010 Int'l Workshop*, Grenoble, June 2010, pp. 1–6.

[33] H. Katti, K. Yadati, M. Kankanhalli, and C. Tat-Seng, "Affective video summarization and story board generation using pupillary dilation and eye gaze," in *Multimedia (ISM 11), 2011 IEEE Int'l Symp.*, 2011, pp. 319–326.

[34] J. J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Multimedia and Expo (ICME 09), IEEE Int'l Conf.*, July 2009, pp. 1436–1439.

[35] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Trans.*, vol. 3, no. 1, pp. 18–31, 2012.

[36] J. A. Russell, "A circumplex model of affect." *J. personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[37] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. of IEEE Int'l Conf. Automatic Face and Gesture Recognition (FG'11), EmoSPACE 2011 - 1st Int'l Workshop Emotion Synthesis, rePresentation, and Analysis in Continuous spacE*, Santa Barbara, CA, USA, March 2011, pp. 827–834.

[38] Y. Cui, J. S. Jin, S. Zhang, S. Luo, and Q. Tian, "Music video affective understanding using feature importance analysis," in *Proc. the ACM Int'l Conf. Image and Video Retrieval*, ser. CIVR '10, NY, USA, 2010, pp. 213–219.

[39] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *Image Processing (ICIP 09), 16th IEEE Int'l Conf.*, 2009, pp. 1853–1856.

[40] A. Hanjalic and L.-Q. Xu, "User-oriented affective video content analysis," in *Content-Based Access of Image and Video Libraries, 2001 IEEE Workshop*, 2001, pp. 50–57.

[41] A. Hanjalic, "Multimodal approach to measuring excitement in video," in *Multimedia and Expo, 2003 Int'l Conf.*, 2003, pp. 289–292.

[42] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features," in *Multimedia and Expo (ICME 08), IEEE Int'l Conf.*, 2008, pp. 1369–1372.

[43] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, "Emotional identity of movies," in *Image Processing (ICIP 09), 16th IEEE Int'l Conf.*, 2009, pp. 1821–1824.

[44] C. H. Chan and G. J. Jones, *An affect-based video retrieval system with open vocabulary querying*. Springer, 2011.

[45] ——, "Affect-based indexing and retrieval of films," in *Proc. 13th Ann. ACM Int'l Conf. Multimedia*, NY, USA, 2005, pp. 427–430.

[46] J. H. French, "Automatic affective video indexing: Sound energy and object motion correlation discovery: Studies in identifying slapstick comedy using low-level video features," in *Southeastcon, 2013 Proc. IEEE*, 2013, pp. 1–6.

[47] S. Moncrieff, C. Dorai, and S. Venkatesh, "Affect computing in film through sound energy dynamics," in *Proc. the 9th ACM int'l Conf. Multimedia*, ser. MULTIMEDIA '01, NY, USA, 2001, pp. 525–527.

[48] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[49] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP J. on Image and Video Processing*, vol. 1, no. 26, pp. 1–10, 2013. [Online]. Available: http://jivp.eurasipjournals.com/content/2013/1/26

[50] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proc. 1st int'l ACM workshop Music information retrieval with user-centered and multimodal strategies*, NY, USA, 2011, pp. 7–12.

[51] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proc. the ninth ACM int'l conf. Multimedia*. ACM, 2001, pp. 203–211.

[52] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Soc. for Eng. Education (ASEE) Zone Conf. Proc.*, 2008, pp. 1–7.

[53] M. Radmard, M. Hadavi, M. M. Nayebi *et al.*, "A new method of voiced/unvoiced classification based on clustering," *J. Signal and Information Processing*, vol. 2, no. 04, p. 336, 2011.

[54] T. Zhang and C.-C. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *Speech and Audio Proc., IEEE Trans.*, vol. 9, no. 4, pp. 441–457, 2001.

[55] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[56] K. R. Scherer and M. R. Zentner, "Emotional effects of music: Production rules," *Music and emotion: Theory and research*, pp. 361–392, 2001.

[57] R. W. Picard, *Affective computing*. MIT press, 2000.

[58] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *J. personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[59] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation." in *INTERSPEECH*, 2010, pp. 785–788.

[60] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[61] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. 16th ACM Int'l Conf. Multimedia*, 2008, pp. 677–680.

[62] P. N. Juslin and J. A. Sloboda, *Music and emotion: Theory and research.* Oxford Univ. Press, 2001.

[63] D. Liu, L. Lu, and H. Zhang, "Automatic mood detection from acoustic music data." in *ISMIR*, 2003.

[64] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. J. Scott, J. A. Speck, and D. Turnbull, "State of the art

report: Music emotion recognition: A state of the art review," in *Proc. ISMIR 10*, 2010, pp. 255–266.

[65] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 40:1–40:30, 2012.

[66] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, 2013.

[67] D. Bordwell, K. Thompson, and J. Ashton, *Film art: an introduction*. McGraw-Hill New York, 1997, vol. 7.

[68] G. M. Smith, *Film structure and the emotion system*. Cambridge Univ. Press, 2007.

[69] C. R. Plantinga and G. M. Smith, *Passionate views: Film, cognition, and emotion*. Johns Hopkins Univ. Pr, 1999.

[70] B. H. Detenber, R. F. Simons, and G. G. Bennett Jr, "Roll em!: The effects of picture motion on emotional responses," *Journal of Broadcasting & Electronic Media*, vol. 42, no. 1, pp. 113–127, 1998.

[71] R. F. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, "Emotion processing in three systems: The medium and the message," *Psychophysiology*, vol. 36, no. 5, pp. 619–627, 1999.

[72] B. Adams, C. Dorai, and S. Venkatesh, "Novel approach to determining tempo and dramatic story sections in motion pictures," in *Image Proc., 2000. Int'l Conf.*, vol. 2, 2000, pp. 283–286.

[73] K. Choroś, "Video shot selection and content-based scene detection for automatic classification of tv sports news," in *Internet–Technical Development and Applications*. Springer, 2009, pp. 73–80.

[74] "The why, the where and the when to move the camera," https://www.hurlbutvisuals.com/blog/2013/04/camera-motion-for-filmmakers/.

[75] "Movie making manual, wikibook," http://en.wikibooks.org/wiki/Movie_Making_Manual/Cinematography/Moving_the_camera.

[76] Z. Herbert, "Sight, sound, motion: Applied media aesthetics," 1999.

[77] K. McLaren, "XIII the development of the cie 1976 (l* a* b*) uniform colour space and colour-difference formula," *J. the Soc. Dyers and Colourists*, vol. 92, no. 9, pp. 338–341, 1976.

[78] G. R. Bradski and J. W. Davis, "Motion segmentation and pose recognition with motion history gradients," *Machine Vision and Applications*, vol. 13, no. 3, pp. 174–184, 2002.

[79] P. Valdez and A. Mehrabian, "Effects of color on emotions." *J. Experimental Psychology: General*, vol. 123, no. 4, p. 394, 1994.

[80] C. Y. Wei, N. Dimitrova, and S.-F. Chang, "Color-mood analysis of films based on syntactic and psychological models," in *Multimedia and Expo, 2004 IEEE Int'l Conf.*, 2004, pp. 831–834.

[81] S. Arifin and P. Y. K. Cheung, "A novel probabilistic approach to modeling the pleasure-arousal-dominance content of the video based on " working memory"," in *Semantic Computing (ICSC 07). Int'l Conf.*, 2007, pp. 147–154.

[82] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[83] J. Wang, B. Li, W. Hu, and O. Wu, "Horror video scene recognition via multiple-instance learning," in *Acoustics, Speech and Signal Processing , 2011 IEEE Int'l Conf.*, 2011, pp. 1325–1328.

[84] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Multimedia and Expo, 2005. ICME 2005. IEEE Int'l Conf.* IEEE, 2005, pp. 4–pp.

[85] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Computer Vision (ICCV), 2013 IEEE Int'l Conf.* IEEE, 2013, pp. 3304–3311.

[86] S. Nie and Q. Ji, "Capturing global and local dynamics for human action recognition," in *22nd Int'l Conf. Pattern Recognition*, 2014, pp. 1946–1951.

[87] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MultiMedia Modeling*, 2014, pp. 303–314.

[88] I. J. Roseman, "Cognitive determinants of emotion: A structural theory." *Rev. of Personality & Social Psychology*, 1984.

[89] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[90] S. Wang, Z. Liu, Y. Zhu, M. He, X. Chen, and Q. Ji, "Implicit video emotion tagging from audiences facial expression," *Multimedia Tools and Applications*, pp. 1–28, 2014.

[91] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Multimedia (ISM 08). 10th IEEE Int'l Symp.*, no. 10455061, Dec. 2008, pp. 228–235.

[92] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging [social sciences]," *Signal Processing Magazine, IEEE*, vol. 26, no. 6, pp. 173–180, 2009.

[93] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[94] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *Int'l J. Psychophysiology*, vol. 61, no. 1, pp. 5–18, 2006.

[95] M. K. Abadi, M. Kia, R. Subramanian, P. Avesani, and N. Sebe, "Decoding affect in videos employing the MEG brain signal," in *Automatic Face and Gesture Recognition (FG 13), 10th IEEE Int'l Conf. and Workshops*, 2013, pp. 1–6.

[96] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, "Highlight detection in movie scenes through inter-users, physiological linkage," in *Social Media Retrieval*. Springer, 2013, pp. 217–237.

[97] J. Fleureau, P. Guillotel, and Q. Huynh-Thu, "Physiological-based affect event detector for entertainment video applications," *Affective Computing, IEEE Trans.*, vol. 3, no. 3, pp. 379–385, July 2012.

[98] S. Koelstra, C. Muhl, and I. Patras, "EEG analysis for implicit tagging of video data," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd Int'l Conf.*, 2009, pp. 1–6.

[99] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *Brain Informatics*, 2010, pp. 89–100.

[100] A. T. Krzywicki, G. He, and B. L. O'Kane, "Analysis of facial thermal variations in response to emotion: eliciting film clips," in *SPIE Defense, Security, and Sensing*, 2009, pp. 734 312–734 312.

[101] A. G. Money and H. Agius, "ELVIS: Entertainment-led video summaries," *ACM Trans. Multimedia Computing Comm. Appl.*, vol. 6, no. 3, pp. 17:1–17:30, 2010.

[102] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proc. 2nd ACM Workshop Multimedia Semantics*, 2008, pp. 32–39.

[103] M. Soleymani and M. Pantic, "Multimedia implicit tagging using EEG signals," in *Multimedia and Expo (ICME 13), 2013 IEEE Int'l Conf.*, 2013, pp. 1–6.

[104] A. F. Smeaton and S. Rothwell, "Biometric responses to music-rich segments in films: The CDVPlex," in *Content-Based Multimedia Indexing, 2009. CBMI'09. Seventh Int'l Workshop*. IEEE, 2009, pp. 162–168.

[105] S. Toyosawa and T. Kawai, "An experience oriented video digesting method using heart activity and its applicable video types," in *Advances in Multimedia Information Processing-PCM 2010*, 2010, pp. 260–271.

[106] M. K. Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe, "User-centric affective video tagging from MEG and peripheral physiological responses," in *Affective Computing and Intelligent Interaction (ACII 13), 2013 Humaine Assoc. Conf.*, Switzerland, Sep. 2013, pp. 582–587.

[107] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proc. ACM Int'l Conf. Image and Video Retrieval*, 2009, pp. 31:1–31:8.

[108] D. McDuff, R. El Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE Int'l Conf. and Workshops*, 2013, pp. 1–7.

[109] W.-T. Peng, C.-H. Chang, W.-T. Chu, W.-J. Huang, C.-N. Chou, W.-Y. Chang, and Y.-P. Hung, "A real-time user interest meter and its applications in home video summarizing," in *Multimedia and Expo (ICME), 2010 IEEE Int'l Conf.* IEEE, 2010, pp. 849–854.

[110] W. T. Peng, W. J. Huang, W. T. Chu, C. N. Chou, W. Y. Chang, C. H. Chang, and Y. P. Hung, "A user experience model for home video summarization," in *Advances Multimedia Modeling*, 2009, pp. 484–495.

[111] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *Multimedia, IEEE Trans.*, vol. 8, no. 3, pp. 500–508, 2006.

[112] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford univ. press, 2007.

[113] W. T. Peng, W. T. Chu, C. H. Chang, C. N. Chou, W. J. Huang, W. Y. Chang, and Y. P. Hung, "Editing by viewing: automatic home video summarization by viewing behavior analysis," *Multimedia, IEEE Trans.*, vol. 13, no. 3, pp. 539–550, 2011.

[114] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose, "Enriching user profiling with affective features for the improvement of a multimodal recommender system," in *Proc. the ACM Int'l Conf. Image and Video Retrieval*, 2009, pp. 29:1–29:8.

[115] Z. Liu, S. Wang, Z. Wang, and Q. Ji, "Implicit video multi-emotion tagging by exploiting multi-expression relations," in *Automatic Face and Gesture Recognition (FG 13), 2013 10th IEEE Int'l Conf. and Workshops*, 2013, pp. 1–6.

[116] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *Affective Computing, IEEE Trans.*, vol. 3, no. 1, pp. 42–55, 2012.

[117] M. Soleymani, "Implicit and automated emotional tagging of videos," Ph.D. dissertation, Univ. of Geneva, Londan, 2011.

[118] I. Arapakis, I. Konstas, and J. M. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," in *Proc. 17th ACM Int'l Conf. Multimedia*, 2009, pp. 461–470.

[119] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," in *Affective Computing and Intelligent Interaction (ACII 13), 2013 Humaine Assoc. Conf.*, Switzerland, Sep. 2013, pp. 13–18.

[120] A. Schaefera, F. Nilsb, X. Sanchezc, and P. Philippotb, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[121] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *the SIGIR 2010 Workshop Crowdsourcing for Search Evaluation*, 2010.

[122] J. Tarvainen, M. Sjoberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2085–2098, 2014.

[123] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *Multimedia, IEEE Trans.*, vol. 12, no. 7, pp. 682–691, 2010.

[124] M. Abadi, R. Subramanian, S. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *Affective Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[125] J. D. Laird and C. Bressler, "The process of emotion experience: a self-perception theory," *Multimedia Systems*, pp. 213–234, 1992.

[126] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2011, vol. 6974, pp. 107–116.

[127] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[128] P. Philippot, "Inducing and assessing differentiated emotion-feeling states in the laboratory," *Cognition and Emotion*, vol. 7, no. 2, pp. 171–193, 1993.

[129] Z. Wang, S. Wang, M. He, Z. Liu, and Q. Ji, "Emotional tagging of videos by exploring multiple emotions' coexistence," in *Automatic Face and Gesture Recognition (FG 13), 2013 10th IEEE Int'l Conf. and Workshops*, 2013, pp. 1–6.

[130] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *Multimedia, IEEE Trans.*, vol. 13, no. 6, pp. 1356–1370, 2011.

[131] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends ® in Machine Learning*, vol. 2, no. 1, pp. 1–127.

[132] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *CVPR 2011*, 2011, pp. 2857–2864.

[133] D. Yu, G. Hinton, N. Morgan, J. T. Chien, and S. Sagayama, "Introduction to the special section on deep learning for speech and language processing," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 20, no. 1, pp. 4–6, 2012.

[134] Z. Wang, S. Lyu, G. Schalk, and Q. Ji, "Learning with target prior," in *Annual Conf. Neural Information Processing Systems (NIPS 12)*, 2012, pp. 2231–2239.

[135] ——, "Deep feature learning using target priors with applications in ECoG signal decoding for BCI," in *Proc. the Twenty-Third int'l joint conf. Artificial Intelligence*. AAAI Press, 2013, pp. 1785–1791.

[136] X. Hu, K. Li, J. Han, X. Hua, L. Guo, and T. Liu, "Bridging the semantic gap via functional brain imaging," in *Multimedia, IEEE Trans.*, 4 2012, pp. 314–325.

[137] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009.

[138] P. Winoto and T. Y. Tang, "The role of user mood in movie recommendations," *Expert Systems Applications*, vol. 37, no. 8, pp. 6086–6092, 2010.

[139] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd Int'l Conf.*, 2009, pp. 1–7.

[140] C. Lithari, C. Frantzidis, C. Papadelis, A. B. Vivas, M. Klados, C. Kourtidou-Papadeli, C. Pappas, A. Ioannides, and P. Bamidis, "Are females more responsive to emotional stimuli? a neurophysiological study across arousal and valence dimensions," *Brain Topography*, vol. 23, no. 1, pp. 27–40, 2010.

[141] R. Srivastava, S. Yan, T. Sim, and S. Roy, "Recognizing emotions of characters in movies," in *Acoustics, Speech and Signal Proc. (ICASSP), 2012 IEEE Int'l Conf.* IEEE, 2012, pp. 993–996.

[142] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *Multimedia, IEEE Trans.*, vol. 16, no. 4, pp. 1075–1089, June 2014.

[143] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. H. Hovy, "Learning whom to trust with mace." in *HLT-NAACL*, 2013, pp. 1120–1130.

Shangfei Wang received the B.S. degree in Electronic Engineering from Anhui University, China, in 1996. She 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 (USTC), China, in 1999 and 2002. From 2004 to 2005, she was a postdoctoral research fellow in Kyushu University, Japan. Between 2011 and 2012, Dr. Wang was a visiting scholar at Rensselaer Polytechnic Institute, USA. She is currently an Associate Professor of School of Computer Science and Technology, USTC. Dr. Wang is an IEEE and ACM member. Her research interests cover computation intelligence, affective computing, and probabilistic graphical models. She has authored or co-authored over 70 publications.

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). 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 IEEE and IAPR.