Highlight Detection in Movie Scenes Through Inter-users, Physiological Linkage

Christophe Chênes, Guillaume Chanel, Mohammad Soleymani, and Thierry Pun

Abstract Automatic summarization techniques facilitate multimedia indexing and access by reducing the content of a given item to its essential parts. However, novel approaches for summarization should be developed since existing methods cannot offer a general and unobtrusive solution. Considering that the consumption of multimedia data is more and more social, we propose to use a physiological index of social interaction, namely, physiological linkage, to determine general highlights of videos. The proposed method detects highlights which are relevant to the majority of viewers without requiring them any conscious effort. Experimental testing has demonstrated the validity of the proposed system which obtained a classification accuracy of up to 78.2%.

1 Introduction

The rapid rate of multimedia creation motivates the development of novel multimedia management and indexing methods. These methods are facing many challenges such as the semantic gap and the complexity of multimedia data. In this context, tagging is an essential step for an effective indexing of multimedia content. In

C. Chênes (✉) • M. Soleymani • T. Pun
Computer Science Department, Computer Vision and Multimedia Laboratory (CVML)

University of Geneva, Battelle Campus, Building A, 7 route de Drize, CH-1227 Carouge, Geneva, Switzerland
e-mail: christophe.chenes@gmail.com; mohammad.soleymani@unige.ch; thierry.pun@unige.ch

G. Chanel
CVML and Swiss Center for Affective Sciences, 7 rue des Battoirs, CH-1205, Geneva, Switzerland
e-mail: guillaume.chanel@unige.ch

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_10, © Springer-Verlag London 2013

addition, since tags are usually given by users, this approach is likely to reduce the semantic gap. Summarization techniques are fundamental to manage large-scale multimedia collections, since they allow to reduce the content to its essential parts. Summarization can thus be considered as a form of tagging with video sequences being "tagged" as a highlight (i.e., a relevant and essential part of the video) or a non-highlight.

Major improvements have been made in the multimedia indexing and retrieval field over the last years thanks to user-generated tags and social networks. Some of these tagging methods have brought a new approach, human-centered tagging, by which the data are tagged by the users. An improvement to these methods is presented in the previous chapter with the use of implicit tagging which aims at reliably extracting tags from nonverbal behaviors of users who are confronted to multimedia data [1]. These new tags, such as emotional keywords, are supposed to be more robust, more useful, and more convenient for content recommendation or retrieval as they are unobtrusive and uncontrollable.

In the perspective of highlight detection in movie scenes, emotions play an essential role. The movie structure, the plot, and the narrative are carefully designed by directors to elicit strong emotions from the spectators. This is the reason why movies have often been used for emotion elicitation studies [2]. The emotional moments of movies can then be considered as strong highlights. As a consequence, the spectators' emotional responses are reliable indicators for highlight detection. Moreover, as emotions enhance the memory [3], highlights corresponding to the spectators' emotional response peaks are potentially the most memorable ones, which is also a desirable property of detected highlights.

Automatic detection of emotions is one of the main goals of affective computing [4]. This field of research is thus of high interest for implicit tagging and emotional highlights detection. Researchers have proposed to use several modalities to automatically detect human emotions [5–7]. These include facial expressions, speech, and also physiological signals. The physiological signals have certain advantages, for example, when measured with the appropriate sensors, physiological signals are difficult to hide or fake. So far, affective computing has mostly focused on the detection of a single person's emotional experience. However, emotions often emerge during social interactions, and it would thus be valuable to switch the unit of analysis from the person to the group. This can, for instance, be achieved by computing physiological linkage which measures the extent to which the physiological signals of two people are dependent of each other [8].

Physiological linkage can occur during any social interaction, and also due to the common interpretation and perception of a stimulus. For example, the spectators, feeling empathy with the movie characters, having similar emotional reactions to the movie content and experiencing social-emotional contagion in the case of multi viewers' show, are expected to have synchronized and similar physiological reactions. It is expected that this form of linkage increases during moments of a movie when most of the spectators are strongly moved and engaged in the movie. Physiological linkage can then be considered as a reliable indicator of highlights when most spectators are emotionally and similarly impacted [9].

This chapter proposes a new approach for highlight detection in videos. The proposed approach is based on the analysis of several spectators' physiological signals. It is thus a social approach that provides a user-independent and content-independent or general (as opposed to personalized and content dependent) summarization of the videos by aggregating the subjective experiences of all spectators. The use of physiological measures also has the advantage of not obstructing the watching of the video in comparison with methods that require users to explicitly provide feedback. One of the central ideas of the proposed work, linked with the social media, is thus to record several spectators' physiological responses and to consider high physiological linkage sequences [10] as highlights. This technique enables us to obtain an accurate and user-independent summary of a given video. It is important to clarify that this method does not require emotion recognition, since it is sufficient to compute linkage directly between spectators' physiological signals. However, it is expected that linkage is a result of the synchronization of spectators' emotional responses.

Therefore, this human-centered method uses the viewers' physiological responses to tag a video sequence either as a highlight (the positive class) or a non-highlight (the negative class). Based on the literature, we expect that movies elicit strong and synchronous emotional responses in spectators that can be detected by measuring their physiological signals. Consequently, the system we propose for highlight detection is based on the following hypothesis: *the moments when viewers' physiological responses are highly linked correspond to the highlight sequences of the scene*.

To answer this hypothesis, this chapter is structured as follows. Firstly, a summary of the work related to the proposed method is given. It focuses on the importance of social considerations for affective tagging in indexing and retrieval processes and also describes physiological linkage. It then thoroughly explains the existing video summarization techniques, both internal and external with a particular focus on the closest techniques to the proposed work. Secondly, the user-independent system developed is described. This includes explanations of the concept, the signal processing, the physiological linkage method applied, the classification methods used, and the stimuli-reaction delay management. Thirdly, the proposed system is evaluated based on the physiological data collected in a previous experiment. In this phase particular attention was given to the collection of a reliable groundtruth. The results validating the proposed approaches are then presented and discussed. Finally, this chapter ends with a summary of the study, the remaining issues, and the future work.

2 Related Work

2.1 *Social and Affective Tagging*

The act of tagging is a well-known and largely applied method which has been used very often to simplify the description of complex data by experts in their domain. With the development of the Web 2.0, this method has become collaborative [11]. The action of associating a tag to a given content is no longer done by a single expert but by a large number of users who are free to use their own terms to describe the content of the data they are rating. When there is a very large amount of data to process, such as on the Web, the collaborative tagging process is the most effective [12]. However, since any user is free to tag with any term, this approach is not fully reliable. Users tend to tag data for personal motivations (selfishness) and social motivations (reputation) [13, 14]. Therefore, applications based on this process suffer from lack of objectivity and reliability. These limitations also demonstrate the importance of social factors in tagging processes.

A solution to improve tagging reliability is to rely on the implicit cues given by multimedia consumers rather than on their explicit evaluation of the multimedia content. This human-centered method is known as implicit tagging [1] and is discussed in the previous chapter. Most of the studies on implicit tagging have focused on affective tags since affect is a highly relevant criterion for multimedia indexing and retrieval based on preferences. Affective tags are also useful to partially bridge the semantic gap. The tags used in this method are linked with spontaneous reactions of the users to the multimedia content they are watching/listening; they can be based on the user's facial expression [15] or physiological reactions [16–18].

Affective implicit tagging is possible thanks to the advancement of affective computing [4]. Affective computing has two major goals:

1. The detection and recognition of emotions: Computers should be able to detect users' emotional changes based on different signals recorded through specific sensors; they should also be able to recognize users' emotions.
2. Computer affective reactions: Computers should be able to synthesize emotions and empathetic reactions and to improve and create a naturalistic human-computer interaction; this branch is motivated by the Turing test [19].

The research in this innovative field has lead to several models for the detection and recognition of emotions. It has also described three main channels for affective sensing: visual, audio, and physiological. Within the human-computer interaction (HCI) development, affective computing applications range from e-health service [20] to video games [21,22]. Independently from the topic, every study in affective computing has helped to better understand the user's experience, including those employing physiological signals [23–25]. Different physiological signals from both the central and the peripheral nervous system have been used for the purpose of emotion assessment [6,26,27]. This includes electroencephalography (EEG), which measures brain electrical potentials; electrodermal activity (EDA), which measures the resistance of the skin (an index of perspiration), blood pressure (BP), heart rate (HR), respiration, and temperature; and electromyography (EMG), which measures the electrical activity originated from muscular activity.

Studies on affective computing, so far, were mostly focused on emotional experience of a single user. However, multimedia is produced and consumed socially. This is, for instance, the case of music and movies which are very often produced by bands and teams that interact during the whole creation phase. This social aspect of creation has, for instance, encouraged the creation of social musical instruments [28]. Furthermore, people enjoy sharing and exchanging media, as well as the experiences they felt during their consumption. Finally, they also get together in theaters and festivals to watch movies and listen to music. There is now clear evidence that the emotional expression of a person can shape the emotions of their peers [29, 30]. The experience of users should thus not be considered independently of each other, and some measures of joint-mediated social interactions and associations should be determined. This is in line with the work done in the field of social signal processing [31] which aims at improving the social abilities and social intelligence [32] of computers. In this context, a measure of social interaction can help to better understand and quantify social processes [33] but can also be used to provide new multimedia tags. In the range of social interactions, we include any social association, affective bond, and affinities that can exist between two people experiencing similar content (e.g., if spectators have similar preferences and thus react in the same way to the movie or if spectators are friends and share a common ground).

2.2 Social Interaction and Physiological Linkage

We use social signals in our daily interactions to send messages and mediate our social behaviors. The signals given from one communicator to another are mainly gesture, posture, facial expression, and voice prosody. However, physiological signals also carry relevant social information since they are modulated by affective reactions [34]. Social effects do not only occur in a classical one-to-one communication; interactions also happen when experiencing shared material, such as music and movies [28]. Participants are socially linked when watching such multimedia content together, they feel the others' presence, and they share similar emotions, such as joy through smiling or laughter and fear through screaming. This social context induces emotional contagion [29] which is the emotional convergence of participants who are influenced by others' reactions. This fact represents the importance of social context on physiological responses since it emphasizes the linkage between participants' reactions. From another perspective, in the case of movies, viewers can feel empathy with a character. Depending on their involvement in the movie, they can experience social presence by identifying themselves with a movie character or imagining themselves in the character's situation [33, 35].

As proposed in [33], physiological linkage is an interesting measure to study mediated social interactions. Physiological linkage has been originally proposed by Gottman [8] and measures to what extent physiological signals of two people depend on each other. The most straightforward method to infer physiological linkage is to compute correlation between two physiological signals. However, several other methods can be used [33]: coherence measures allow to determine if signals oscillate at the same frequency by switching from the temporal to the frequency domain; Granger causality and similar methods based on linear auto-regressive models account for nonindependence of time series samples; and methods from nonlinear dynamical systems, such as synchronization measures [36], are able to deal with nonlinear signals.

Physiological linkage has been shown to be related to several social processes. In [37] the authors observed a higher level of physiological linkage during conflicting interactions compared to low-intensity nonconflicting interactions of married couples. Levenson et al. [38] also demonstrated that physiological linkage is related to the accuracy of rating others' feelings which can be considered as an indicator of empathy. More recently Henning et al. [39, 40] have analyzed physiological linkage in the context of collaborative processes. They discovered that linkage is associated with team performance [39] when playing on a collaborative video game. However, these results were contradicted by a later study [40] that concluded physiological linkage was inversely proportional to self-reported team productivity, quality of communication, and ability to work together. While the studies cited above focus on the use of peripheral physiological signals, it is also possible to compute linkage and synchronization on brain signals. For instance, it has been demonstrated that inter-person brain signal synchronization occurs during imitation of the other [41]. Concerning movies, it has been shown in [42] that synchronization can also occur during natural visualization of films and it is argued that interbrain synchronization can be understood as supporting the coordination of actions and the common understanding of the environment [43]. Finally, the analysis of synchronicity and linkage is not limited to physiological signals and can also be used to measure synchronous social behaviors [28] between musicians. All these results clearly demonstrate that physiological linkage is associated to several types of measurements and situations that involve social interactions and associations.

2.3 Automatic Highlight Detection in Videos

The automation of highlight detection or extraction from video is a far-reaching subject. With the exponential growth of video media, it becomes essential to have some algorithms able to summarize video content. The aim of video abstraction is, in general, to put together the highlights of the video. The first step of video summarization techniques is then the detection of highlight sequences which are considered as relevant by the majority of viewers. The second step of a complete summarization method includes the cut of the highlight sequences from the video stream, respecting the video structure (scenes, clusters, shots) [44, 45] and the reconstruction of the summary with the selected highlights.

The vast quantity of videos induces a huge variation of genres. The differences between genres can be significant and, most important, the definition of highlights can be extremely diverging. For example, a soccer match video summary will be mainly composed by goal actions, while the highlight of a drama scene will be a wave of sadness. It is then obvious that the goal of a general method raises great difficulties.

In the current literature, two categories of video abstraction methods are reported [46, 47] as well as the combination of these two categories:

Internal summarization techniques are based on the analysis of low-level features present within the video stream, such as color or speech. Many interesting techniques have been developed with promising results. For example, a method based on the grass proportion in the video, the slow-motion effect, and the cinematic of the scene was implemented for soccer match summarization to detect goals, slow motions, and referee actions [48]. Its global correct classification rate reaches 89%. The recall of the method is impressive, more than 90%, but the precision is about 40%. However, internal techniques face a certain number of remaining challenges, in particular the semantic gap and their highly domain-specific design which makes the goal of a general method utopian, though they appear to be very popular and effective techniques for the sport videos [48–50] due to the structures of sport games and the interferences of the spectators.

External summarization techniques are based on features entirely external to the video, such as user's description of the video. The information used by these techniques to summarize the video can be of two types:

1. User-based, sourced directly from the user, such as from the facial expression of the user [51]
2. Contextual, sourced from the context of the user but not the user himself, such as the video recording GPS position [52]

These techniques try to summarize videos with a higher level of abstraction. They reflect better the comprehension of the user and thus reduce the semantic gap. In spite of this significant advantage over the internal summarization techniques, such systems have been rarely implemented [46, 47].

In survey written by Money et al. [46], three key external techniques are reviewed. The first one is contextual and unobtrusive as it does not require the user to give any information explicitly. It consists fixed-position video cameras and integrated pressure-based floor sensors tracking the location of user activity in the home during the shooting stage. The video can then be summarized according to the characters' positions in the house and their footsteps analysis. The two other reported studies are both obtrusive since based on manual annotations of the users [53, 54]. These descriptions are detailed and domain specific, in this case for baseball and soccer. Moreover, two out of the three techniques reported

Table 1 ELVIS results compared to random

	Comedy		Horror/comedy		Horror	
	Random	ELVIS	Random	ELVIS	Random	ELVIS
Mean on 20 users	30.37%	45.96%	31.34%	43.77%	28.38%	41.74%

produce personal summary and are therefore user-dependant. Although external summarization techniques might reach a general method, existing approaches are too costly to the end user.

Another existing external technique is the detection of personal highlights through facial expression [51]. This method tracks motion vector of 12 points on the participant's face. An experiment on ten participants was conducted. The participants watched eight video clips of different genres and reported highlight annotations at the end of the clip. The best precision result is 40%, and the authors reported that the best point on the face varies for each participant. Consequently, this method cannot be extended to a user-independent technique.

In the perspective of what has been done, the analysis of user's physiological responses for video summarization, a new user-based external summarization technique, can dramatically reduce the user effort and improve the external summarization techniques. This is an almost unexplored branch. The most and only advanced study, to our best knowledge, is *ELVIS* [47]. It is a complete system of personal video summarization based on five physiological signals (electrodermal activity, heart rate, blood volume pulse, respiration rate, and respiration amplitude). It is motivated almost by the same reasons than this project : video content elicits strong physiological responses from the user, highlight sequences elicit stronger responses, and these responses can be detected by recording physiological signals. The authors explore whether user's physiological responses can serve as new information to produce personalized video summaries. The result is a user personalized video summary corresponding to the favorite segments of the user. The signals from a single user are processed to obtain high and low values and are then combined to form a physiological significance index. Highlights are identified as the most significant segments. The system is precisely evaluated thanks to a large-scale experiment conducted over 60 participants. Not only physiological signals were recorded during these sessions but participants also reported what was the highlight sequences for them. This allowed a thorough analysis of their system (cf. Table 1). *ELVIS* has showed the usability of users' physiological responses as information for video summarization. However, they used single-user signals in order to compute personalized summary and therefore cannot propose a user-independent approach.

As a summary, most of the reported internal techniques focus on sports highlights which reduce their interest since they are not extendable to other genres of videos. One of the disadvantages of existing summarization techniques is their lack of users' self-reported highlights for ground truth construction. The evaluation of the techniques is then not based on the opinion of media consumers but on the evaluations of a unique experimenter. Finally, the few external summarization techniques reported are designed to produce personalized summarization for a single user. This lack of generalization over both the type of videos (sports, drama, horror, comedy, etc.) and the population is the main drawback of existing techniques. In order to achieve a user-independent and content-independent video summarization technique, it is necessary to compute the summary based on the information gathered from several users and to evaluate such a system on a strong reference obtained through users' self-reports.

3 System Definition

The new user- and content-independent, unobtrusive external video summarization technique proposed in this chapter is defined in the following section.

3.1 Concept

The system aims to provide a novel highlight detection method in order to compute user-independent summaries of videos. The major challenges are the difficulty to reach a general method – that is, a method which detects sequences considered as relevant for a majority of viewers and content independent – and the difficulty to have an unobtrusive approach, that is, an approach which will not require any effort to the end user. Considering all the facts reported in the previous section about tagging, the right way forward seems to be a human-centered implicit tagging. This approach is the most likely to bridge the semantic gap since it relies on the user's comprehension of the multimedia content.

With the evolution of affective computing, recording of emotional reactions through physiological signals allows to detect user's affective changes. Using modern sensors, these parameters can be obtained without any conscious effort from the user. This represents therefore an unobtrusive technique able to provide reliable tags. Since emotional reactions are common to everyone, they can lead to user independent tags. The system is designed as a user-based external summarization technique which tags every instant of the video as highlight or non-highlight on the basis of the participants' physiological responses.

Although emotions make sense to everyone, not everyone has exactly the same reactions to movies; thus, to guarantee a user-independent method, it is necessary to combine the physiological signals of several participants together to obtain the user-independent highlight sequences. This is the main difference with the existing summarization techniques based on single-user physiological responses, which produces personalized summaries [47].

The concept of the system is to compute inter-users, physiological linkage to detect user-independent highlight sequences without any effort required from users. With this design, the system is likely to detect user-independent highlights.

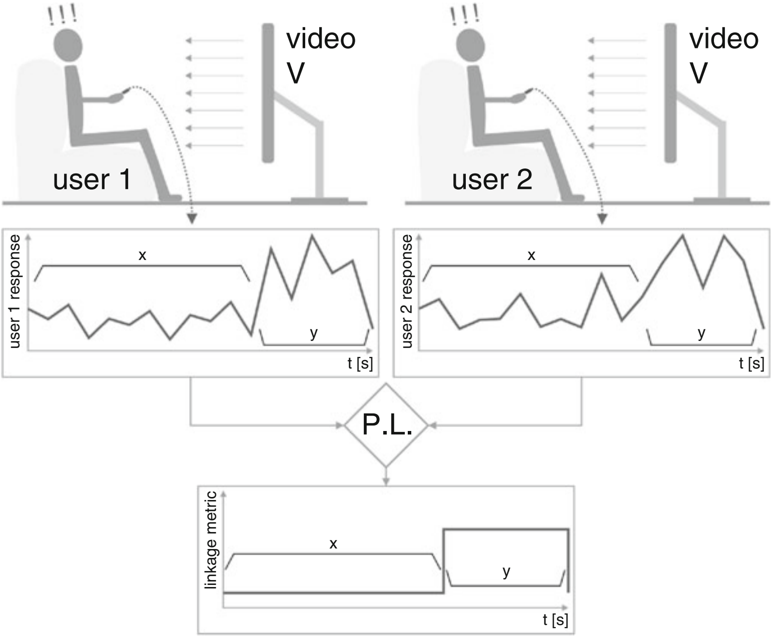

Fig. 1 Physiological linkage (P.L.) system concept

Moreover, with its social approach, it allows to detect social interaction, such as emotional contagion, which should enhance the generalizability of the algorithm. As the emotions improve the memory [3], the highlights detected with this system should be the most memorable ones. The method belongs to the category of external summarization techniques but only concerns the detection of highlights and does not include the summary reconstruction.

3.2 Physiological Linkage

The system concept is based on the linkage between several participants' emotional responses. These responses are represented by participants' physiological signals which are recorded during the viewing of the videos and stored as time series. Figure 1 presents the scheme of the physiological linkage concept with two participants (*user 1* and *user 2*). One of their physiological signals is recorded; while they are watching *video V*, the corresponding time series is presented for each participant. These two signals are then given to the physiological linkage (*P.L.*) unit which computes the physiological linkage and gives a metric vector of the linkage as a result. Highlight sequences can then be extracted from this metric as the highest values. On the figure, two periods can be identified. The *x* period corresponds to the signals randomly oscillating around their mean and corresponding to a low physiological linkage (non-highlight). The *y* period corresponds to an event in *video V* inducing some coupling in the physiological activity and resulting in high physiological linkage (highlight). In this example, there is only one highlight sequence, but the system is designed to detect as many highlights as the scene contains.

The physiological linkage computation is realized with a sliding-window linear correlation between every possible pairs of participants. This metric was chosen for its simplicity and efficiency and because it is used in other works on physiological linkage [33, 39, 40]. This continuous linkage metric is computed on a window time frame which is shifted of half the window length at each iteration. When more than two participants' physiological signals are available, it results in as many linkage metric vectors as number of pairs. These vectors form then a matrix of linkage metrics, and the final vector is computed as the mean of all values at each time interval.

3.3 Sequence Classification

Once the global linkage metric vector is obtained for a video, the highlights are extracted by classification using the best limit on the metric. Therefore, several highlights can be detected in a scene. This detection is performed by a support vector machine (SVM) classifier in its quadratic mode, which was previously trained with the datasets presented in the *system evaluation* section. Two approaches are experimented: the first one uses only a single signal to extract the highlight and the second one combines multi-signals to extract the highlights. This second method aims to obtain better results by increasing the system dimensions and by taking advantage of physiological signal combination.

4 System Evaluation

The proposed system was tested on a database of physiological signals and its results are thoroughly analyzed to determine whether the user-independent, content-independent, and unobtrusive system goal is reached

4.1 Physiological Signals Recording and Preprocessing

The physiological signals used to test the proposed system were recorded in an experiment previously conducted for the purpose of emotional implicit tagging [55]. This dataset is composed of 64 scenes extracted from eight movies at the rate of eight scenes by movie. A subset of 26 scenes out of the 64 was created

Table 2 Physiological signal preprocessing

EMG	High-pass 10 Hz, absolute value, running average window of size 0.5 s, logarithm
BPM (from BVP)	Low-pass 5 Hz, beat detection, beat correction
EDA	Low-pass 3 Hz, logarithm
Skin temperature	Low-pass 1 Hz

by experimentators for this work. The selection is distributed among four major genres:

- Action: `Saving Private Ryan and Kill Bill, Vol.1`
- Drama: `Hotel Rwanda and The Pianist`
- Comedy: `Mr. Bean's Holiday and Love Actually`
- Horror: `28 Days Later and Ringu`

A selection of various movie genres is necessary to test the generality of the system. Each scene lasts about 2 min and, based on the experimentators' judgment, contains an emotional event.

The physiological signals were recorded for eight healthy participants, three females and five males from 22 to 40 years old. Six physiological signals were recorded, but only five are used in this work:

- *Electromyogram (EMG) from right zygomaticus major*: measure of the activation of this muscle, involved in smile and laughter
- *Electromyogram (EMG) from right frontalis*: measure of the activation of this muscle, involved in attention and surprise
- *Blood volume pulse (BVP)*, using a plethysmograph: measure of the relative change of blood pressure on the top of the thumb
- *Electrodermal activity (EDA)*: measure of the skin resistance between the index and the middle fingers
- *Skin temperature*: measure of the skin temperature with a temperature sensor placed on the top of the little finger

The preprocessing of each kind of signal was done according to the guidelines found in the literature [56–58] (see Table 2). For all the signal a high-pass filter was applied to remove drifts. Some signals were also corrected by computing their logarithm in order to normalize their range across participants. The envelope of the EMG signals was computed by taking the absolute value of the filtered EMG and applying a running average window. Heart beats were detected from the BVP signals by identification of local maximums.

4.2 Highlights in Scenes Reference

The demonstration of the system generality goes through a thorough evaluation which requires a highlight ground truth in the used scenes. This information is

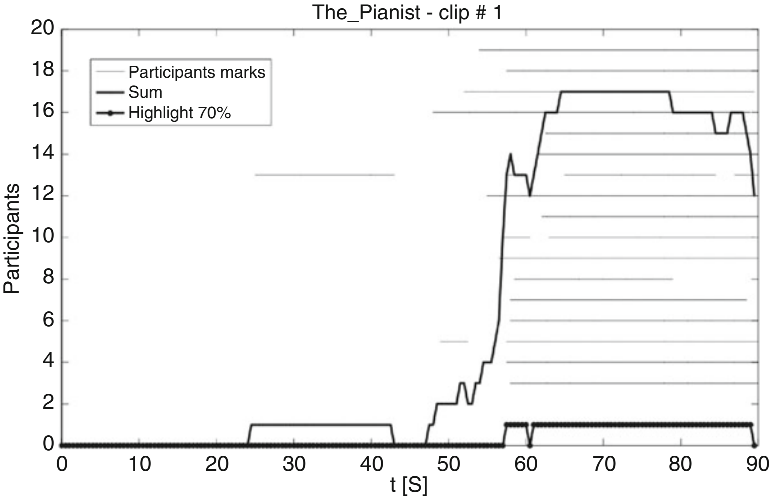

Fig. 2 Highlights ground truth example

compulsory to train and test the classifier and to compute statistical results. It was obtained during another experiment, conducted on 18 participants who did not participate in the physiological signal recording experiment. The large number of participants enabled to obtain the user-independent definition of highlights in the selected scenes.

Participants had to watch a scene once entirely and then were asked to mark the sequences they judged as highlights through a simple, original interface. The whole process was repeatedly done for each scene. The scenes were randomly presented to each participant. From these experiments, a vector containing the sum of highlights marked at each instant by the 18 participants was obtained for each scene. Each instant corresponds to each 0.5 s of the scene. An instant was finally judged as a highlight when 70% of the participants marked it. Figure 2 shows an example of result for a scene of `The Pianist` for which participants agreed on the highlight. The marked moment corresponds to a bombing near the main character after a peaceful sequence of piano. The horizontal thin lines represent each participant's marks, the curve is the sum, and the thick line is the final highlight definition used for the system evaluation.

As participants did not agree about highlight annotations for each scene, the coherence among participants' annotations was computed with the Fleiss' kappa. This was done to remove scenes on which the participants did not clearly agreed on the presence of highlights. This metric, ranging from 0 to 1, informs on the coherence, and thus the generality, of the highlights marked by the participants. Figure 3 shows the coherence distribution for the 26 scenes. It appeared clearly that the distribution was bimodal which conducted to the creation of a subset of scenes. This second dataset contains 13 scenes for which the kappa was over the median of the distribution and the highlights are general. The system was evaluated on both the 13 scene and the 26 scene datasets.

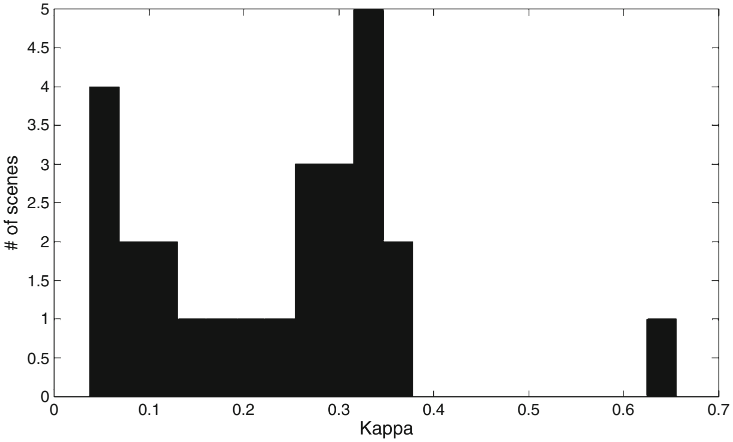

Fig. 3 Coherence distribution among the participants' annotations

4.3 Specifications for System Parameters

The physiological linkage unit of the system is based on the sliding-window linear correlation, which has two main parameters: the window length and the overlap between two consecutive windows. Considering that the scenes last about 2 min a window length of 10 s with a time lag of half the window was chosen.

The use of a 10-s window with a 5-s shift is optimal for the detection of events but can artificially create a delay. In addition, some signals have a long reaction time, especially the skin temperature. These two facts induce a delay between the stimuli – that is, the event in the video – and the reaction, that is, the change of the physiological linkage index, which has an impact on the system. Figure 4 presents an example of clear delay between the stimuli – the plain curve, and the reaction, the dashed curve. This figure also demonstrates that highlights can be detected in the physiological response. Indeed, the two highlight sequences with ground truth reported on the figure correspond in the video (a `28 Days Later` scene), respectively, to a very short movement of a dead character after a distressing period of calm which is surprising but not extremely relevant and to the assault of the main character by a zombie, which is an emotionally intense moment and therefore induces a strong physiological linkage. Several delays were then tested depending on the signal to realign the curves. Their choice is discussed in the next section.

4.4 Results

The system was evaluated on two setups, a single-signal analysis in which only one kind of signal at a time was used and a multi-signal analysis in which all signals

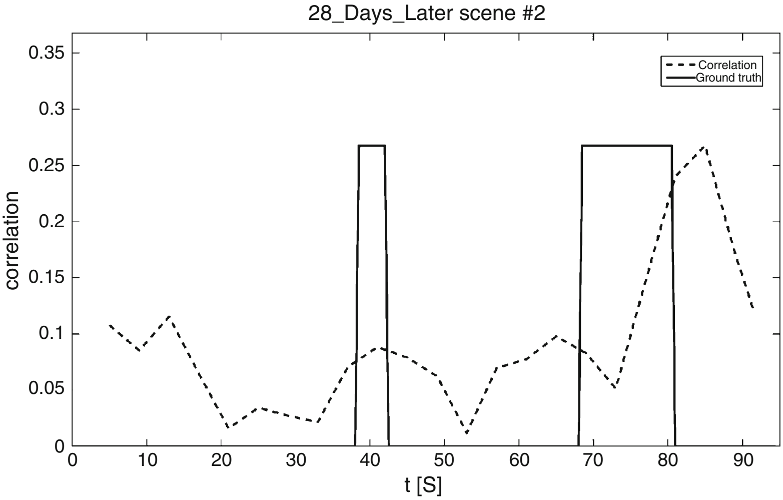

Fig. 4 Stimuli-reaction delay example, high value for ground truth represents highlight sequence

Table 3 Highlight detection results for the *skin temperature* signal with the delay influence

Delay	−10 [s]	*−8 [s]*	−6 [s]	−4 [s]	−2 [s]	0 [s]
ROC area	0.692	*0.701*	0.692	0.674	0.647	0.619
F1-score	0.547	0.549	*0.567*	0.549	0.549	0.515
Accuracy	0.766	*0.775*	0.760	0.745	0.723	0.720

were combined together. The results were analyzed through the *receiver operating characteristic (ROC)* curve, characterized by the area under the curve, the accuracy, and the *precision-recall (P-R)*, characterized by the F1-score.

The single-signal analysis allowed to identify the best signal for the highlight detection process as the *skin temperature* signal and also proved the significant, positive influence of the stimuli-reaction delay on the results. Highlight detection results for the skin temperature are presented in Table 3 in which the delay influence is clearly visible and the best delay results are displayed in bold. The *ROC* and *P-R* curves are displayed on Fig. 5 for the *skin temperature* signal with the best delay of 8 s. In this configuration, the system returned a correct classification rate of *77%*. In comparison, a random classifier obtains 50% and a classifier returning always the majority class (in this case non-highlight) obtains 72% of accuracy. The first and foremost observation is thus that the proposed system is able to identify user- and content- independent highlights in video scenes.

The second analysis which combined all the signals together aimed to improve the results by increasing the number of dimensions, which could lead to a better separation between the two classes, and by disclosing physiological dependencies among the signals. The classification was performed by a *support vector machine (SVM)* using its quadratic mode. The training/testing phases were conducted using

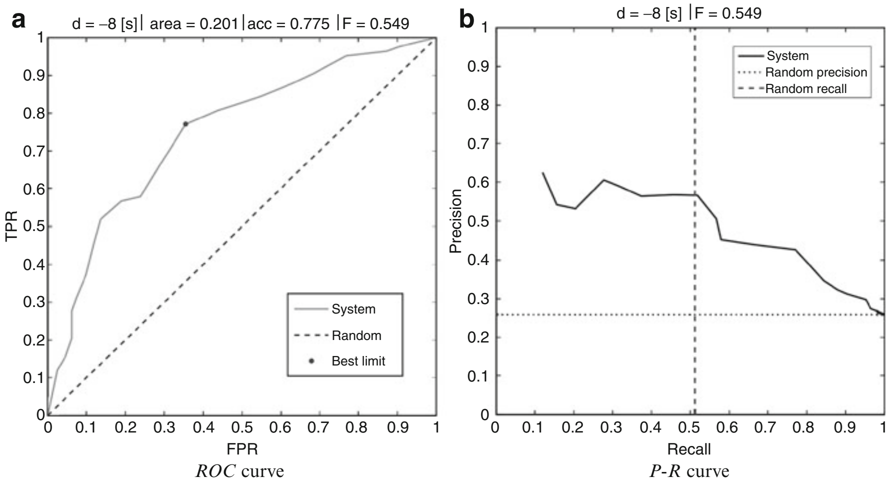

Fig. 5 *ROC* and *P-R* curves for the *skin temperature* signal with the 8-s delay (*d*). On the *top* of the figures, the area under the *ROC* curve (*area*), the accuracy (*acc*), and the F1-score (*F*) are displayed. (**a**) *ROC* curve and (**b**) *P-R* curve

a *leave-one-out* cross-validation technique. This second method did improve the results, obtaining *78.2%* of correct classification, but had only a limited influence. Nevertheless, it was proven that the system works even better by increasing the number of physiological signals involved in the highlight detection process.

What is surprising about the single-signal result is that it is the *skin temperature* signal which returned the best correct classification rate, whereas this signal was not expected to be the most informative for the detection process because of its very slow response time. On the contrary, the *EDA* signal was considered as the most promising signal because of its correlation with felt arousal and finally returned results inferior to the *skin temperature* signal, that is, 75% of correct classification. Even more surprisingly, the two *EMG* signals turned out to be totally uninformative; their results were not even better than what a random classifier would obtain. What the single-signal approach teaches is that using only one signal, a satisfactory, general, highlights detection method is obtained. The original system which is proposed in this chapter has then been proved to be workable.

Even though it has been shown that the multi-signal approach improved the results in comparison to the single-signal approach, does this approach really improve the system? Considering that four additional signals were necessary to obtain 1.2% of better classification, it may represent too much work for such a small benefit. If the design was applied on a larger scale, such as at home through the Web for online video summarization, the single-signal method, especially with the use of the *skin temperature* signal, is clearly advantageous since it only requires a cheap, simple sensor to record it.

As the viewing of a movie should elicit smiles and frowns, reactions driven by *EMGs* should be informative. This lack of emotional reaction demonstrated through the obtained results could be explained by the context of the physiological recording experiment: the participants were alone in front of a monitor in a laboratory environment, and short scenes were randomly submitted to their viewing. This is extremely different from a multi-person ecological situation in a theater or at home. The emotional contagion occurring when watching a film surrounded by several people in a theater was missing, and the participants could feel not very at ease in the laboratory and thus could limit their facial expressions. In addition, as there was no narrative context and the scenes duration were short, the participants were not allowed to be in the mood of the scene, and the randomly submitted scenes from various genres could induce residual affective states: a participant's emotional reaction to a funny scene differs depending on whether the previous scene was a horror scene or a comedy scene [2]. Consequently, the context of the physiological recording experiment appears to be a key point.

The system provides better results than the closest work, *ELVIS*, which only obtains between 40 and 45% of correct classification on three genres of videos [47] and than the work using facial expressions which obtains an F-score of only 0.150 for highlight detection [59]. In addition, though many internal summarization techniques obtain better result in their own context, for example, 86.4% of accuracy for sports highlights correctly detected in [49], they are certainly not capable to detect highlights for other kinds of videos. However, the proposed system still suffers from two drawbacks: too many false positives were retrieved and the delays were chosen without cross-validation. These issues must be resolved through future work.

5 Conclusion

This project aimed to provide a user-independent and content-independent video highlight detection method based on inter-users, physiological linkage. Confidence was granted to this approach on the basis of its novel use of unobtrusive implicit human-centered tagging able to take advantage of the social interaction occurring between spectators and their common emotional interpretation of a movie. The system uses correlation on inter-users, physiological signals to compute the physiological linkage during the scene. It then tags the video sequences as highlight or non-highlight depending on the physiological linkage metric. The very interesting results obtained, 77% of correct classification with only the *skin temperature* signal and 78.2% of correct classification with all the signals combined together, demonstrate that the system works better than existing external summarization techniques and is context independent.

However, several limitations have been identified, such as the physiological recording context, the low precision, and the empirical choice of the delays. The following future work is thus necessary. Firstly, the ecological and physiological recording should be conducted on complete movies with several participants at the same time in an actual cinema. Of course, this raises many difficulties, but the results and the hypothetical improvements obtained will be extremely interesting. Secondly, the system precision as long as the system results could benefit from the use of another physiological linkage unit. Since the correlation is a very simple, linear method, a more complex, nonlinear mathematical tool, such as the synchronization likelihood, is likely to improve the results. Finally, the delay choice must be asserted through cross-validation.

The implemented system led to results which enabled the verification of the research hypothesis: the moments when viewers' physiological responses are highly linked correspond to highlight sequences of the video. The feasibility of a user-independent and content-independent, unobtrusive method for highlight detection in videos using inter-users, physiological linkage has been demonstrated. Even if a great amount of work should still be accomplished to reach a complete summarization tool, it represents a progress in external video summarization method, which could be, in the near future, widely used.

References

1. Pantic, M., Vinciarelli, A.: Implicit human centered tagging. IEEE Signal Process. Mag. **26**, 11 (2009)
2. Gross, J.J., Levenson, R.W.: Emotion elicitation using films. Cognit. Emot. **9**(1), 87–108 (1995)
3. Hamann, S.: Cognitive and neural mechanisms of emotional memory. Trends Cognit. Sci. **5**(9), 394–400 (2001)
4. Picard, R.W.: Affective computing. M.I.T media laboratory perceptual computing section technical report, 321 (1995)
5. Zeng, Z.H., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
6. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. Int. J. Hum. Comput. Stud. **67**(8), 607–627 (2009)
7. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proc. IEEE **91**(9), 1370–1390 (2003)
8. Gottman, J.M.: Detecting cyclicity in social-interaction. Psychol. Bull. **86**(2), 338–348 (1979)
9. Money, A.G., Agius, H.: Analysing user physiological responses for affective video summarisation. Displays **30**(2), 59–70 (2009)
10. Levenson, R.W., Gottman, J.M.: Marital interaction: physiological linkage and affective exchange. J. Personal. Soc. Psychol. **45**(3), 587–597 (1983)
11. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**, 198–208 (2006)
12. Golder, S., Huberman, B.A.: Huberman: Usage Patterns of Collaborative Tagging Systems. J. Inf. Sci. **32**(2), 198–208 (2006)
13. Nov, O., Naaman, M., Ye, C.: What drives content tagging: the case of photos on flickr. In: Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pp. 1097–1100. ACM, New York(2008)

14. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07, pp. 971–980. ACM, New York (2007)
15. Jiao, J., Pantic, M.: Implicit image tagging via facial information. In: Proceedings of the 2nd International Workshop on Social Signal Processing, SSPW '10, pp. 59–64. ACM, New York (2010)
16. Koelstra, S., Mühl, C., Patras, I.: Eeg analysis for implicit tagging of video data. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 27–32. IEEE, Los Alamitos (2009)
17. Yazdani, A., Lee, J.S., Ebrahimi, T.: Implicit emotional tagging of multimedia using eeg signals and brain computer interface. In: Proceedings of the First SIGMM Workshop on Social Media, WSM '09, pp. 81–88. ACM, New York (2009)
18. Soleymani, M., Chanel, G., Kierkels, J.J.M., Pun, T.: Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Proceedings of the Tenth IEEE International Symposium on Multimedia, pp. 228–235, 15–17 Dec (2008)
19. Turing, A.M.: Computing Machinery and Intelligence, pp. 11–35. MIT, Cambridge (1995)
20. Lisetti, C., Lerouge, C.: Affective computing and tele-home health. In: Proceedings of the 37th Hawaii International Conference on System Sciences, pp. 148–155, Orlando, FL, USA, 5–8 Jan. (2004)
21. Paiva, A., Prada, R., Chaves, R., Vala, M., Bullock, A., Andersson, G., Höök, K.: Towards tangibility in gameplay: building a tangible affective interface for a computer game. In: Proceedings of the 5th International Conference on Multimodal Interfaces, ICMI '03, pp. 60–67. ACM, New York (2003)
22. Chanel, G., Rebetez, C., Betrancourt, M., Pun, T.: Emotion assessment from physiological signals for adaptation of games difficulty. IEEE Trans. Syst. Man Cybern. Part A **41**(6), 1052–1063 (2011)
23. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1175–1191 (2001)
24. Kim, K., Bang, S., Kim, S.: Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. **42**, 419–427 (2004). doi:10.1007/BF02344719
25. Scheirer, J., Fernandez, R., Klein, J., Picard, R.W.: Frustrating the user on purpose: a step toward building an affective computer. Interact. Comput. **14**(2), 93–118 (2002)
26. Katsis, C.D., Katertsidis, N., Ganiatras, G., Fotiadis, D.I.: Toward emotion recognition in car racing drivers: a biosignal processing approach. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(3), 502–512 (2008)
27. Rani, P., Liu, C., Sarkar, N.: An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal. Appl. **9**, 58–69 (2006)
28. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Trans. Multimed. **12**(6), 576–590 (2010)
29. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Emotional contagion. Curr. Dir. Psychol. Sci. **2**, 96–100 (1993)
30. Jakobs, E., Fischer, A.H., Manstead, A.S.R.: Emotional experience as a function of social context: the role of the other. J. Nonverbal Behav. **21**(2), 103–130 (1997)
31. Pentland, A.: Social signal processing [exploratory dsp]. Signal Process. Mag. IEEE **24**(4), 108–111 (2007)
32. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009). Visual and multimodal analysis of human spontaneous behaviour
33. Ekman, I., Chanel, G., Kivikangas, J.M., Salminen, M., Järvelä, S., Ravaja, N.: Social interaction in games: measuring physiological linkage and social presence. Simul. Gaming **43**(3), 321–338 (2012)

34. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous-system activity distinguishes among emotions. Science **221**(4616), 1208–1210 (1983)
35. Coplan, A.: Catching characters' emotions: emotional contagion responses to narrative fiction film. Film Stud. **8**, 26–38 (2006)
36. Stam, C.J.: Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clin. Neurophysiol. **116**(10), 2266–2301 (2005)
37. Levenson, R.W., Gottman, J.M.: Marital interaction: physiological linkage and affective exchange. J. Personal. Soc. Psychol. **45**(3), 587–597 (1983)
38. Levenson, R.W., Ruef, A.M.: Empathy: a physiological substrate. J. Personal. Soc. Psychol. **63**(2), 234–246 (1992)
39. Henning, R.A., Boucsein, W., Gil, M.C.: Social-physiological compliance as a determinant of team performance. Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol. **40**(3), 221–232 (2001)
40. Henning, R.A., Armstead, A.G., Ferris, J.K.: Social psychophysiological compliance in a four-person research team. Appl. Ergon. **40**(6), 1004–1010 (2009)
41. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L.: Inter-brain synchronization during social interaction. PloS one **5**(8), 1–10 (2010)
42. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. Science (New York) **303**(5664), 1634–1640 (2004)
43. Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., Keysers, C.: Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in Cognit. Sci. **16**, 114–121 (2012)
44. Ngo, C.-W., Ma, Y.-F., Zhang, H.-J.: Automatic video summarization by graph modeling. IEEE Int. Conf. Comput. Vis. **1**, 104 (2003)
45. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Trans. Multimed. **7**, 1097–1105 (2005)
46. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
47. Money, A.G., Agius, H.: Elvis: entertainment-led video summaries. ACM Trans. Multimed. Comput. Commun. Appl. **6**, 17:1–17:30 (2010)
48. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Trans. Image Process. **12**(7), 796–807 (2003)
49. Wang, J., Xu, C., Chng, E., Tian, Q.: Sports highlight detection from keyword sequences using hmm. In: Proceedings of the IEEE ICME, Taipei, pp. 27–30 (2004)
50. Li, J., Wang, T., Hu, W., Sun, M., Zhang, Y.: Soccer highlight detection using two-dependence bayesian network. IEEE Int. Conf. Multimed. Expo **0**, 1625–1628 (2006)
51. Joho, H., Staiano, J., Sebe, N., Jose, J.M.: Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. Multimed. Tools Appl. **51**, 505–523 (2011)
52. Aizawa, K., Tancharoen, D., Kawasaki, S., Yamasaki, T.: Efficient retrieval of life log based on context and content. In: Proceedings of the the 1st ACM workshop on Continuous Archival and Retrieval of Personal Experiences, CARPE'04, pp. 22–31. ACM, New York (2004)
53. Jaimes, A., Jaimes, R., Echigo, T., Teraguchi, M., Satoh, F.: Learning personalized video highlights from detailed mpeg-7 metadata. In: MPEG-7 Metadata, in Proceedings of the ICIP, p. 2002. IEEE, Piscataway (2002)
54. Takahashi, Y., Nitta, N., Babaguchi, N.: Video summarization for large sports video archives. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2005, Amsterdam, pp. 1170–1173 (2005)
55. Soleymani, M., Chanel, G., Kierkels, J.J.M., Pun, T.: Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In: Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia, ISM '08, pp. 228–235. IEEE, Washington, DC (2008)
56. Fridlund, A.J., Cacioppo, J.T.: Guidelines for human electromyographic research. Psychophysiology **23**(5), 567–589 (1986)

57. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Handbook of Psychophysiology. Cambridge University Press, Cambridge/New York (2000)
58. Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal response system. In: Cacioppo, J.T., Tassinary, L.G. (eds.) Principles of Psychophysiology: Physical, Social and Inferential Elements, pp. 295–324. Cambridge University Press, Cambridge (1990)
59. Joho, H., Jose, J.M., Valenti, R., Sebe, N.: Exploiting facial expressions for affective video summarisation. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09, pp. 31:1–31:8. ACM, New York (2009)