Corpus Development for Affective Video Indexing

Mohammad Soleymani, *Member, IEEE*, Martha Larson, *Member, IEEE*, Thierry Pun, *Member, IEEE*, and Alan Hanjalic, *Senior Member, IEEE*

***Abstract*—Affective video indexing is the area of research that develops techniques to automatically generate descriptions of video content that encode the emotional reactions which the video content evokes in viewers. This paper provides a set of corpus development guidelines based on state-of-the-art practice intended to support researchers in this field. Affective descriptions can be used for video search and browsing systems offering users affective perspectives. The paper is motivated by the observation that affective video indexing has yet to fully profit from the standard corpora (data sets) that have benefited conventional forms of video indexing. Affective video indexing faces unique challenges, since viewer-reported affective reactions are difficult to assess. Moreover affect assessment efforts must be carefully designed in order to both cover the types of affective responses that video content evokes in viewers and also capture the stable and consistent aspects of these responses. We first present background information on affect and multimedia and related work on affective multimedia indexing, including existing corpora. Three dimensions emerge as critical for affective video corpora, and form the basis for our proposed guidelines: the context of viewer response, personal variation among viewers, and the effectiveness and efficiency of corpus creation. Finally, we present examples of three recent corpora and discuss how these corpora make progressive steps towards fulfilling the guidelines.**

***Index Terms*—Benchmarks, content analysis, emotional characterization, multimedia, videos.**

I. Introduction

VIDEO indexing is the process of analyzing video content in order to extract a representation that is specific enough to characterize the uniqueness of the content and, at the same time, is abstract enough to capture useful similarities with other video content. Research and development in the area of video indexing falls under the larger domain of multimedia content analysis, which includes the theories, algorithms and systems that extract or infer descriptors which encode characteristics of multimedia content. These descriptors take a variety of forms, ranging from machine interpretable indexing features to metadata labels in the form of textual words or phrases that can also be interpreted directly by humans (e.g., [1], [2], [3]). The common function of such descriptors is to represent video content in a way that makes possible systems that give users better access to multimedia content. In particular, here, we are interested in video indexing techniques that will be used for video search engines and other systems that support browsing video collections or otherwise represent to users the contents of a video stream.

Conventionally, video indexing has focused on describing videos in terms of the content that humans identify as being explicitly depicted in their visual channel. Much attention has been devoted to developing algorithms that detect visual concepts in video that are related to events, objects, people, scenes, and locations [4]. Such concepts can be considered the 'literal' content of a video. The meaning or the value of a particular video for a viewer clearly goes far beyond its literal content, however. Videos can also be characterized in terms of how they influence viewers' emotions, i.e., their affective impact on viewers. Affective viewer response refers to the intensity and type of emotion that is evoked in a viewer while watching a video. The potential of affective indexing to contribute to the automatic creation of descriptions that are useful for video search engines is widely acknowledged. However, much research in the area of video indexing remains focused on literal descriptions of video and affective video indexing has yet to reach its full potential.

An important factor contributing to the success of visual concept detection and other literal approaches to indexing video is the existence of standardized corpora (data sets). These corpora are made available to the research community, often within the framework of a benchmarking initiative, and can be used by researchers to evaluate the algorithms that they develop. For example, detection of visual concepts in video have been a primary focus for the largest multimedia benchmarking efforts, most notably TRECVid [5], [6]. Similar large, high-quality data sets, used at the community level in benchmarking initiatives, have yet to be developed for affective video indexing.

This paper takes the position that corpora have a key role to play in supporting the research work that is necessary in order to allow affective video indexing to reach its full potential. The main contribution of this paper is a set of 'affective video indexing corpus development guidelines' that arise from a discussion of the state of the art and an analysis of the limitations of existing data sets. The guidelines are organized along three dimensions that are identified as critical for the process of corpus

Manuscript received November 09, 2012; revised April 23, 2013, August 25, 2013, and April 23, 2013; accepted December 06, 2013. Date of publication February 10, 2014; date of current version May 13, 2014. The work of M. Soleymani was supported by the European Research Area under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The work of M. Larson and A. Hanjalic was supported in part by the European Community's Seventh Framework Program under grant agreement no. 287704 (CUbRIK). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sen-Ching Cheung.

M. Soleymani is with the Intelligent Behaviour Understanding Group (iBUG), Imperial College London, London SW7 2AZ, U.K. (e-mail: m.soleymani@imperial.ac.uk).

M. Larson and A. Hanjalic are with the Multimedia Information Retrieval Lab, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: m.a.larson@tudelft.nl; a.hanjalic@tudelft.nl).

T. Pun is with the Computer Vision and Multimedia Laboratory, University of Geneva, Carouge(GE) CH-1227, Switzerland (e-mail: thierry.pun@unige.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2305573

development for affective video indexing: the context of viewer response, personal variation among viewers, and the effectiveness and efficiency of the process of collecting viewer-reported affective reactions.

A. The Rise of Affective Video Indexing

The affective video indexing paradigm assumes that users' focus in selecting multimedia content involves a strong affective component and that a multimedia information system, e.g., a video search engine, must be able to take feelings, emotion and mood into account. Recently, the importance of affect in people's information seeking behavior has been recognized, as witnessed by work in the area of conventional text information retrieval, such as [7], [8]. In parallel, awareness of the potential of affective indexing for multimedia information retrieval has also increased.

Affective video retrieval was first discussed in the mid-1990's by Rosalind Picard as an application of affective computing [9]. Affective video indexing is well summarized by her statement, "Although affective annotations, like content annotations, will not be universal, they will still help reduce time searching for the 'right scene'." [9] (p. 11). When it was first introduced, the proposal that affect could provide an effective means to organize video was not immediately widely accepted. However, a decade later, the idea had matured in form and established its status as a new paradigm within multimedia information retrieval community [10].

The importance of affect is now widely accepted by researchers, as reflected by [11], a survey of multimedia information retrieval, which states that "On a fundamental level, the notion of user satisfaction is inherently emotional." (p. 3). The current paper is motivated by our conviction that the availability of large, high-quality corpora for the evaluation of affective video indexing will support the multimedia research community in turning its awareness of the importance of affective video indexing into tangible and significant advancement of the state of the art.

B. The Challenge of Affective Video Indexing

The central challenge faced by affective video indexing lies in the difference between descriptions that refer to the affective impact of videos (e.g., "uplifting") and descriptions that refer to the literal content of the video (e.g., "sunrise"). In the case of descriptions of literal content, viewers can quickly and consistently assess whether a description is relevant for a given video, e.g., whether or not a given visual concept is depicted in the video. In making this judgment, they rely on cognitive processing combined with general world knowledge. The judgment is considered to be objective because it can be easily reproduced by consulting a group of viewers, largely independently of the viewers' backgrounds.

Characterizing a video with respect to its affective impact on videos is less clear cut. Information on affect can be gathered by asking viewers to report their emotional response upon watching the video. Affective response is considered to be subjective, since only the subject experiencing the response (i.e., the viewer) is in a position of authority to assess or confirm a particular response. It is tempting to conclude that subjectivity (i.e., the fact that no observer other than the viewer has access to direct knowledge of the viewer's affective response) makes the problem of predicting affective response to a video hopelessly ill defined. Indeed, the affective response evoked in a viewer while watching the video is personal in that it can, and does, differ from person to person. It is also contextual, since it varies when the context in which the video is watched or the underlying mood or physical state of the viewer changes.

However, although it is not clear cut, affective response is far from arbitrary. In many cases, affective impact will be quite consistent and there will be a high level of agreement in affective response across viewers. The challenge of affective indexing for video involves how to identify those aspects of video that trigger emotional reactions across viewers that are stable enough that they can be robustly predicted.

The stability of affective impact is most clearly illustrated in the case of film. Filmmakers are highly skilled in evoking specific emotions in their audience. The high-level of inter-subjective agreement concerning the connotative aspects of film has been studied and used as the basis for an automatic indexing system by [12]. Connotation is that dimension of interpretation that goes beyond literal meaning, and, as such, encompasses a large affective component. Today's video search engines index large quantities of video on the Web. For online videos, it is not possible to apply *a priori* assumptions about the techniques and conventions used in formal film to trigger emotions. However, it is possible to anticipate that there will be a component of viewer response that is grounded in modalities of emotional reaction shared in the audience or arising from common interpretation conventions.

This paper takes the standpoint that by isolating and emphasizing aspects of video for which human judges display a relatively high level of agreement, corpora for the evaluation of affective indexing can be created that can make a contribution to advancing the state of the art comparable to the contribution made by benchmarks that focus on literal descriptions of video content.

C. The Contribution of Corpus Development

Corpus development contributes to advancing the state of the art of multimedia technology by making possible standardized evaluation. Only when a standard data set and ground truth are used, is it possible to directly and fairly compare alternative algorithms. Comparison and reproducibility help to drive forward the state of the art: when researchers know how their algorithms perform with respect to the state of the art, they can better direct their efforts to surpass it and more quickly abandon less promising lines of investigation. Benchmarks and standard tasks/data sets help to eliminate redundancy by enabling direct comparison between algorithms across research sites, increasing the efficiency of the research community by allowing resources to be shared between sites and providing a framework in which researchers can interact in a mixture of collaboration and competition that is stimulating and productive. The impact of the TRECVid evaluation for video has been large and is well documented [13]. However, as mentioned above, TRECVid focuses on literal approaches to video indexing, i.e.,

content explicitly depicted in the visual channel. Corpus development is key to allowing affective video indexing to achieve similar impact.

Corpus development does not strive to promote one particular variety of affective video indexing, but rather if numerous, well-designed multimedia corpora were available, they would contribute in many different ways. Here, we provide some examples of the range of applications in which affective video indexing, retrieval and browsing has been used. In [14], highlights were extracted from baseball programs and in [2] an adaptive approach to sports video highlight detection was proposed and studied in detail for the case of soccer. The usefulness of such affective video indexing techniques is witnessed by the fact that Mitsubishi has already released two products taking advantage of the highlight detection for sport events in Japan [15]. Retrieval of movie clips using multimedia content features and user-assigned keywords was investigated by [3] and [16]. Laughter events have been successfully deployed in videos for navigation [17]. The examples illustrate the spread of application areas that stand to benefit if large, high-quality corpora can be developed in made available to the research community.

Current data sets used to evaluate individual theories and algorithms in affective content analysis are typically limited in size and scope. The limitations are imposed because of the relative difficulty of collecting affective responses from many viewers. These limitations also reduce variability in the elicited affective responses of test users, which facilitates manual annotation and results interpretation, but may ultimately be too narrow for the resulting algorithms to be used in practical situations.

Recently, however, technological developments have provided means for developing a new generation of corpora. Online systems make it possible to ask large numbers of viewers to watch videos and provide information on their affective response. Additionally, the rise of crowdsourcing and large crowdsourcing platforms such as Amazon Mechanical Turk (www.mturk.com) make it possible to more easily recruit large numbers of annotators with a representative spread of backgrounds. Corpora in existence today do not, as yet, fully exploit these resources. The corpus development guidelines set out in this paper aim to encourage the effective use of the new opportunities offered by the Web and by crowdsourcing platforms.

In this study, we set our focus on corpora used to study emotion evoked by video for the purpose of affective video indexing. It is important to note that video corpora are also developed in order to study human emotion directly. Such corpora are developed for the general goal of studying emotion. For example, videos of people laughing would be used to study expressions of positive emotion. Research involving such corpora has been treated elsewhere in the literature [18], [19], [20]. Here, we are interested in the emotion of people watching the video, which is not necessarily the same as the emotion of people appearing in the video. For example, videos of people laughing can either evoke a positive or negative emotion in viewers, depending on their perceptions of the person who is laughing.

A concise discussion of the importance of data sets for affect recognition technologies, emphasizing the pressing need for new corpora of emotion-eliciting films, is included in [21]. In order to cover the full scope of affective computing, multimedia corpora designed to elicit the complete spectrum of human emotional response are necessary. Such corpora should not, *a priori*, be assumed to be useful for the purposes that we address here, namely, affective video indexing. As discussed further below, the types and distributions of emotion triggers found in video may be different than those occurring in more general contexts. However, the message of [21], that large amounts of data must be collected in order to carry out research in the area of affect, is a key point that is common with our own work. Further, like our work, [21] identifies crowdsourcing as providing a highly promising new opportunity to gather emotional response data from human subjects.

II. Emotion in Response to Multimedia

In this section, we provide a specification of the key concepts of affect and multimedia that are used in this paper and cover the relevant related work. We start out by presenting a clear definition of affective viewer response to multimedia as it is applied for affective video indexing and we continue to discuss this definition with respect to the larger field of research on human emotion.

A. Affective Viewer Response in Context of Affective Video Indexing

Affective viewer response, as already mentioned above, is the intensity and type of emotion that is evoked in a viewer while watching a video. Emotions are complex phenomena with affective, cognitive, conative and physiological components [22]. The affective component is the subjective experience conventionally connected with feelings. The cognitive component is the perception and evaluation of the emotional situation. The conative component is the expression of affect, including facial expressions, body gesture, and any other action that has a preparatory function for an emotional situation. The physiological component regulates physiological responses in reaction to the emotional situation, for example, increasing perspiration during a fearful experience. When studying emotion evoked in viewers in response to multimedia, it is important to take the complexity of emotion into account rather than expecting emotion to manifest itself along a single dimension only.

Clearly, it is not useful to draw an overly sharp delineation between the study of viewer response for the purpose of affective video indexing and for other purposes. Some affective video indexing corpora will contain selected video content that allows researchers to focus specifically on video content that triggers particular emotional responses. However, careful attention to the difference between corpora for affective video indexing and other video corpora used to study emotions from different perspectives helps to ensure the suitability of the affective corpus for its particular purpose. In the next subsection, we move on to a more general discussion of research on human emotion, maintaining emphasis on that aspects particularly important for affective viewer response to video.

B. Research on Human Emotion

Because of the complexity of emotion, a good entry point into research on human emotion is a clear picture of what emotion is

not. In particular, in understanding the nature of emotional response, the terms "mood" and "emotion" should be differentiated. The point is a particularly important one, since these terms are sometimes used interchangeably in the literature despite the clear formal distinction between their definitions. Mood is a diffused affective state that is long, slow moving and not tied to a specific object or elicitor whereas emotions can occur in short moments with higher intensities [23].

One of the most well-known and widely-accepted theories that explains the development of emotional experience is appraisal theory. According to this theory, cognitive judgment about, or appraisal of, a situation is a key factor in the emergence of emotions [24], [25], [26]. According to Orthony, Clore and Collins (OCC) [25], emotions are experienced following a scenario comprising a series of phases. First, there is a perception of an event, object or an action. Then, there is an evaluation of the event, object or action according to personal wishes or norms. Finally, perception and evaluation result in a specific emotion or emotions arising. During the appraisal process that gives rise to an emotional experience in response to multimedia content, viewers examine events, situations and objects with respect to novelty, pleasantness, goal, attainability, copability, and compatibility with their norms. Then, the viewers' perceptions induce specific emotions, which changes their physiological responses, motor actions, and feelings.

Emotional processes can be divided into different categories. Here, we mention three processes that apply not only in the general case of emotional response, but also in the specific case of viewer affective response to video: *emotion induction*, *emotional contagion* and *empathic sympathy* [22]. An example of *emotion induction* is when in a TV show a politician's comment makes the viewers angry while the politician himself is not angry. The angry response from the viewers is due to their perception of the situation according to their goals and values. *Emotional contagion* occurs when the viewer only experiences emotion that is expressed in the video. For example, the induced joy as a result of sitcom laughter falls into this category. In the empathic category, the situation or event does not affect the viewer directly, but rather the viewer reproduces the appraisal steps of the characters who are depicted in the video. The empathic reaction is described as *symmetric co-emotion* in cases in which the viewer has positive feelings about the character and *asymmetric co-emotion* in cases in which the viewer has negative feeling about the character [27].

Empathy is a complex phenomenon with both cognitive and affective components. Affective empathy is the primitive response involved in sympathizing with other individuals. On the other hand, cognitive empathy is the intellectual understanding of other people and the rational reconstruction of their feelings [28], [22]. Zillman developed an affective disposition theory for narrative plot [29], [27]. According to this theory, empathic emotions originate with the observation of the actors by viewers. First, a character's actions are morally judged by the viewer and the judgment results in a positive or negative perception of the character. Then, depending on whether the viewer approves or disapproves of the character, the viewer sympathizes either empathically or counter-empathetically. The intensity of the affective response to a video depends on how much viewers identify themselves with the protagonists and to what extent they suspend their own identities [29].

In general, it is challenging to go from appraisal theory and the sources of affective empathy to a technique that analyzes the content to predict viewers' emotions. However, some characteristics of video are quite indicative of affective response. Much video will capture not only action, but the audience of that action. This audience might be the spectators of a sports event or certain characters in a film, who watch and react to events. The reaction of these in-video observers (e.g., laughter or cheering) can provide important clues to how viewers will react to the video. Further, the literature has identified the most important emotion inducing components of movies as being music and narrative structures [22]. Music is clearly instantiated at the signal level, and structure can also to a certain extent be extracted (e.g., the quick shot changes of a chase scene that will correspond to abrupt changes in the visual constitution of a scene). In pursuit of such regularities, researchers have undertaken to develop techniques for affective video content analysis, which we will return to discuss in Section IV.

III. Background and Existing Techniques

This section discuses emotional representations and existing tools that have been developed to collect annotations in the form of these representations are introduced.

A. Emotional Representations

There are different emotional representations including, discrete, and continuous models. Discrete representations of emotions, and their theoretical underpinnings, were originally inspired by the representation scheme of Darwin, who considered emotion important for survival. Discrete representations presuppose the existence of the certain number of basic and universal emotions [23], [30]. Some of the most widely-known research on basic emotions was carried out by Ekman [30], whose work supported the universality of facial expressions of emotions. There is currently no answer to the question of how many different emotions there are, but most lists in use contain 6-14 emotions [23]. The lists of emotions used by psychologists are generally utilitarian by nature. Utilitarian emotions are those that are helpful for adapting to the world and make an important contribution to our well-being [23]. It is important to note that utilitarian emotions may not be the most important emotions for the purpose of affective video indexing. Another type of emotion represented with discrete categories is aesthetic emotions [23], emotions arising in conjunction with the appreciation of beauty or quality (including, "admiration", "ecstasy" and "fascination".)

When human affective response is conceptualized as a discrete set of categories, a particular challenge arises. The challenge involves ensuring that the categories are interpreted similarly across different situations. A readily observable difficulty in explicitly defining the scope and coverage of individual affective categories is cross-lingual variation. Such variation arises since emotion words often do not have exact translations into different languages, e.g., there is no word in Polish that corresponds exactly in meaning to the English word, "disgust" [31]. Scherer [23] takes a pragmatic step towards addressing this issue by proposing a mapping between words or word stems, as used by subjects in free choice emotional

reports, to a set of 36 affect categories. Being limited to the English language, this mapping does not, of course, address the cross-lingual issue, but it makes an important contribution to the comparability of emotion reports and emotion studies.

The challenge of ensuring that emotion categories receive a consistent interpretation contributed to the motivation for the development of dimensional approaches. Wundt [32] was the first to propose a dimensional representation for emotions. Recent tools that have been developed, e.g., by [33], discussed in more detail below, use dimensional approaches to minimize the effect of differences in interpretations of discrete categories and to verify inter-participant consistency, e.g., Bradley and Lang [33]. Dimensional theories of emotion are grounded in idea the that emotions can be represented in a continuous space and that discrete emotions are basically folk-psychological concepts that can be identified with points in this space [34].

Dimensional representations used by psychologists often represent emotions in an n-dimensional space (generally 2- or 3-dimensional). The most well-known example of such a space arises is the 3D valence-arousal-dominance or Pleasure-Arousal-Dominance (PAD) space [35]. This space arises from cognitive theory and is widely used for studying affect and multimedia—we ourselves make use of it for corpus development, as discussed later in the paper. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited. The dominance scale ranges from submissive (or "without control") to dominant (or "in control, empowered"). Fontaine *et al.* [36] proposed adding predictability dimension to PAD dimensions. Predictability level describes to what extent the sequence of events is predictable or surprising for a person.

An advantage of using dimensional representations of emotion is that when people are asked to describe their emotions, they are often better at positioning content in comparison to a reference point (i.e., this video was more exciting than the previous one) compared to the situation where they are asked to provide an absolute score [37]. Methods related to those used by [38] are good examples of approaches that make use of relative annotation.

It is important to note that dimensional representations of emotion should not be considered to supersede discrete representations. Rather, both have an important contribution to make to facilitating consistent representation of emotional response in the face of variability introduced by different subjects, different triggers, different contexts and also the passage of time. The importance of the way in which people talk about their own emotions should not be underestimated. Although psychologists strive to develop representation systems transcending folk-psychological concepts, human language remains an important tool for gaining insight on human emotion. The robustness, reproducibility and relevance of natural language characterizations of emotions is assured via the very process which gave rise to the development of human language to support inter-human communication. The critical factor guiding the development of human language expressing emotion can be considered to be the need for the effective and reliable communication of emotions between humans within the context of a shared understanding of a common emotional space. Recent work has been devoted to the creation of computational models of the conceptual meaning of words that humans used to describe emotion [39]. The models are able to map between emotion vocabularies of different sizes in different languages and also to differentiate between nuances of meaning in large emotion word vocabularies.

A particularly important consideration in the area of affective video indexing is that the emotional representation used must provide a good fit with the types of emotion that are expected to occur in response to triggers in video content. In other words, in order to understand the emotional categories or the dimensions that are most appropriate for assessing viewers' emotion, the video content and the context in which it is being watched must be carefully considered. Dimensions can be also identified based on the application. For example, the arousal of a viewer while watching a live sports event might be substantially different than the arousal of a viewer watching a replay of the same sports match.

Finally, in gathering emotional annotations, co-occurring emotions should be also considered. There are cases of co-occurring emotions from different poles of the spectrum or an emotion that co-occurs with a feeling of ambivalence. For example, some music listeners when asked to report their emotions for positive and negative affect separately, rated both as high [40]. This might be a result of music listeners' preference for sad stimuli due to their underlying negative mood [41].

B. Emotional Self-Reporting Methods

Understanding the "true", underlying emotion that was felt by a participant during an experiment has been always a challenge for psychologists. Multiple emotional self-reporting methods have been created and used so far [42], [43], [23], [33], [44]. Emotional self-reporting can be done either in free-response or forced-choice formats. In the free-response format, the experiment participants are free to express their emotions by words. In the forced-choice format, participants are asked to answer specific questions to indicate their emotions. Forced-choice self-reports in affective experiments use both discrete and dimensional approaches. Discrete-emotion self-reporting tools have been developed that can be used to ask participants to report their emotions with emotional words on nominal and ordinal scales. Dimensional approaches of emotional self-reporting are based on bipolar dimensions of emotions. Emotions can be reported along each dimension using ordinal or continuous scales [33]. Here, we discuss in more detail some popular self-reporting methods which have been used for psychological and human computer interaction research.

Russell [45] introduced the "circumplex model" of affect for emotion representation. In his model, eight emotions; namely, "arousal", "excitement", "pleasure", "contentment", "sleepiness", "depression", "misery" and "distress" are positioned on a circle surrounding a two dimensional activation, pleasure-displeasure space. Starting form these eight categories, 28 emotional keywords were positioned on this circumplex, based on the results of a user study. The advantage of this circumplex over either discrete or dimensional models is that all the emotions can be mapped on the circumplex using only the angle. In this way, all emotions are presented on a circular and one dimensional model.

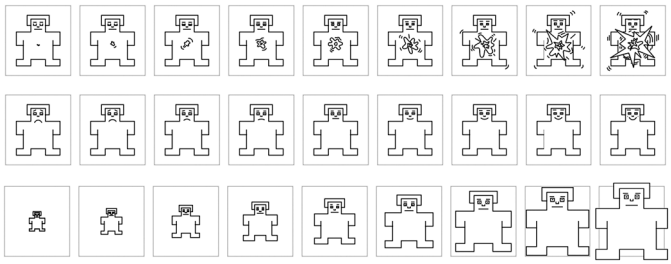

Fig. 1. Self Assessment Manikins. From top to bottom the manikins express different levels of arousal, valence, and dominance.

The Self Assessment Manikin (SAM) is one of the most well-known emotional self-reporting tools. It consists of manikins expressing emotions. The emotions vary along three different dimensions; namely, arousal, valence, and dominance [33]. The SAM Manikins are shown in Fig. 1. Experiment participants can choose the manikin that best portrays their emotion. This method does not require the verbalization of emotions and the manikins are understandable without further explanation. For these reasons, the SAM tool is language independent. The second advantage of the SAM tool is that it can be directly used in measuring the multiple dimensions of emotions. A limitation of SAM is that subjects are unable to express co-occurring emotions with this tool.

The "Positive and Negative Schedule" (PANAS) [46] permits self-reporting 10 positive and 10 negative affects on a five-point scale. An expanded version of PANAS, the "Positive and Negative Schedule—Expanded Form" (PANAS-X), was developed later. PANAS-X provides the possibility of reporting 11 discrete emotion groups on a five-point scale [47]. PANAS is made to report affective states and can be used to report both moods and emotions. PANAS-X includes 60 emotional words and takes on average 10 minutes for an experimental participant to complete [47]. The time needed to answer the PANAS questionnaire makes it too difficult to use in the experiments with limited time and multiple stimuli.

Scherer [23] positioned 20 emotions around a circle to combine both dimensional and discrete emotional approaches, and in this way created the Geneva Emotion Wheel. For each emotion around the wheel, five circles whose size increases from the center outwards are displayed. The size of the circles is an indicator of the intensity of felt emotion (see Fig. 2). In an experiment, participants can pick, from the list of 20 emotions, up to two emotions that were the closest to their experience and report the intensities of the emotions with the size of the marked circles. In case no emotion is felt, a user can mark the upper half circle in the hub of the wheel. If a different emotion is felt by a user, it can be indicated in the lower half circle. The emotions are sorted on the circle such that, high-control emotions are on the top and low-control emotions are at the bottom and the horizontal axis, which is not explicitly visible on the wheel, represents valence or pleasantness.

PrEmo is an alternative non-verbal emotion reporting tool to report emotions in response to product design [42]. It overcomes the problem of reporting co-occurring emotions by making use of animated characters expressing emotions. PrEmo consists of 14 animated characters expressing different emotions and it is, for this reason, language independent. Users can assign a score, at three levels, to one or more characters that they identify as relevant to their emotional response.

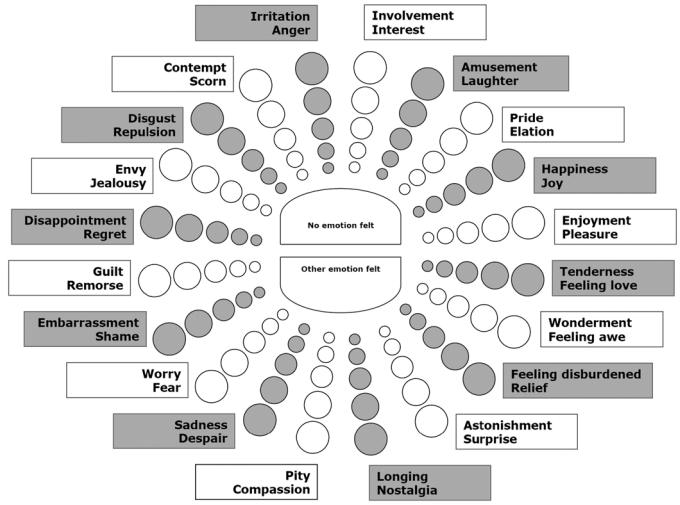

Fig. 2. Subjects can indicate their emotion on the Geneva Emotion Wheel [23] by selecting the corresponding circle.

C. Video Affective Annotation Tools

Among existing self-reporting tools, few have been designed specifically for the affective annotation of video. Villon developed an annotation tool with which a user can drag and drop videos onto the valence-arousal plane [38]. This tool presents the possibility of comparing the ratings given to different videos and enables an experiment participant to rate a video relative to the ratings given other videos. The tool enables users to take their previous reports into account while annotating a new video.

Feeltrace was developed to annotate the intrinsic emotion in videos [48]. This tool was originally designed to annotate the emotions that are expressed by people who are depicted in videos (e.g., in talk shows), including acted facial expressions or gestures [49]. Although this tool provides the possibility of continuous annotation, it is not a particularly appropriate tool for emotional self-reporting, because it is difficult for tool users to both concentrate on the video and at the same time changes in their own emotions. Inspired by Feeltrace, an online annotation interface was created for crowdsourcing platforms [50]. Although the interface was used for the emotional annotations of music, it can be directly used for continuous emotional annotation of videos.

An online video affective annotation tool has been developed by Soleymani *et al.* [51]. With their annotation tool, a experiment participant can self-report emotions after watching a given video clip by means of SAM manikins and emotional keywords from a selected list in a drop down menu. This tool is used in the development of our web-based corpus presented in Section VI-B.

IV. Affective Video Indexing

In this section, we provide an overview of affective indexing and discuss the corpora that have been developed in previous

work. In particular, we discuss the shortcomings of existing corpora, which are used as a basis to develop a set of guidelines for the design and development of future corpora.

Affective video content analysis involves estimating the affective response elicited in viewers by the content. Motivated by work in the area of film, researchers have extracted content features, such as audio energy and color histograms from the video signal and used machine learning techniques to infer which emotion would be felt by an average viewer. They have considered different goals and applications for their algorithms, from video summarization to personalized content delivery.

A. Existing Corpora

In general, affective video corpora are developed with specific goals. Three common goals are: first, emotion elicitation or mood regulation in psychological experiments; second, emotional characterization of videos using content for video indexing or highlighting and third, recognition of the intrinsic emotions in the videos, e.g., detecting the emotions which were expressed by people in the videos [18], [52]. Although all three involve emotion, it is critical to avoid mixing these three different research tracks and the goals behind them. For example, movie excerpts that are most likely to elicit strong emotions are chosen for emotion elicitation in the case of the first goal. In contrast, for the second goal, an exclusive focus on strongly emotional excerpts is not appropriate for emotional characterization. Emotional characterization should be able to deal with the full spectrum of emotions in videos, from neutral videos to mixed and strong emotions.

Rottenberg *et al.* [53] created an emotional video data set for psychological emotion elicitation studies. The excerpts, which were about 1–10 minutes long, were either extracted from famous commercial movies or from non-commercial videos that were used in emotional research, e.g., an amputation surgery video. First, they formed a set of excerpts with different targeted emotions; namely, amusement, anger, disgust, fear, neutral, sadness and surprise. They evaluated the excerpts based on "intensity" and "discreteness". The "intensity" of an excerpt means whether a video received high mean report on the target emotion in comparison to other videos. The "discreteness" refers to the extent to which the target emotion was felt more intensely in comparison to all non-targeted emotions. "Discreteness" was calculated using the ratings a video received on the target emotion in comparison to the other emotions. Ultimately, the data set that was formed consisted of 13 videos, from under a minute to up to eight minutes long, for emotion elicitation studies.

In a more recent study, Schaefer *et al.* [46] created a larger data set from movie excerpts to induce emotions. The study went beyond discrete basic emotions and developed a corpus including 15 mixed feelings in addition to six discrete emotions; namely, "anger", "disgust", "sadness", "fear", "amusement", and "tenderness". 364 participants annotated their database using three questionnaires.

Wang and Cheong [54] created and annotated a data set consisting of 36 full length Hollywood movies having 2040 scenes. Three annotators watched the movies and reported their emotions specifying Ekman basic emotion labels [55] for every scene. Only 14% of the scenes received double labels and the rest only received single emotional labels from their three annotators.

Hanjalic and Xu [56] used excerpts without annotations from the movies "Saving Private Ryan" and "Jurassic Park 3" and two soccer matches in their study. Irie *et al.* [57] used 206 selected emotional scenes out of 24 movies. A total of 16 students annotated these scenes using eight Plutchik basic emotions: "joy", "acceptance", "fear", "surprise", "sadness", "disgust", "anger", and "anticipation" [58]. The annotators first watched the videos and then reported how much they felt each of these emotions on seven-point scale. The emotional labels were assigned to the selected scenes only if more than 75% of annotators agreed on them, otherwise the neutral label was assigned to the movie scene. M. Xu *et al.* [59] used selected scenes from eight movies containing 6201 shots totaling 720 minutes. The videos were manually labeled by five emotions: "fear", "anger", "happiness", "sadness", and "neutral", spanning the arousal dimension in three levels and valence in two levels.

Soleymani *et al.* [60] used 21 full length commercially produced movies. One annotator annotated the movies continuously using an annotation tool which was recording the coordinates of mouse on valence and arousal plane on every click. The annotator reported his emotion at every moment he felt a different emotion while watching the movies.

Teixeira *et al.* [61] used selected excerpts from 24 movies. They first segmented the movies into short clips ($\mathrm{M} = 112\mathrm{s}$), and showed them to 16 participants. Participants rated the movies using SAM Manikins [33] on a seven-point scale; 346 clips, 10 h 26 min in total, were chosen to span arousal, valence, dominance space.

Demarty *et al.* [62] created a benchmark consisting of 18 Hollywood movies for violence detection. Although the movies are not annotated directly with emotional terms, we include this example here since depiction of violence elicits a variety of strong emotions. The data set is annotated by seven annotators on shot level.

B. Open Issues with Existing Corpora

We end this section with a summary of the limitations of currently existing corpora in the form of a catalogue of open issues.

Users' Mood and Context: We have pointed out that the same viewer can experience different emotions in response to the same stimulus depending on the context. The importance of the influence of mood for viewer affective response to video is relatively uncontroversial. For example, it is not strange or surprising when someone remarks, "I am not in the mood to watch that movie today." Our survey has revealed that the assumptions and methodology adopted by existing work is inconsistent with the importance of context for affective reactions. Researchers often fail to emphasize controlling the context and conditions in which annotations are collected from users, or disregard the issue of context entirely when designing experiments and developing data sets. Ignoring or suppressing context introduces risk into affective video indexing research: a system that does not take into account the high degree of variability that characterizes naturally occurring contexts in which video is consumed may not be able to respond appropriately to user needs in real-world situations. As we will discuss further in Section VI, there are a

wide variety of contextual dimensions with a significant effect on the emotions that viewers feel in response to a video, including time of the day, temperature, mood and social context.

User Variability: Beyond contextual factors, most of the existing research on affective video characterization has assumed reactions to be homogeneous across viewers, e.g., [54]. In some cases, the assumption of a single, obvious affective reaction from viewers is so strong, that affective video analysis is carried out, without collecting any user annotations at all, e.g., [56]. In most cases, however, assuming that everyone will react in the same way when watching a particular video is a strongly limiting assumption, that contradicts our intuition that the subjective nature of affect includes a strongly individual dimension. Corpora that allow both the personal and general dimensions of affective reactions to be explored have greater potential in helping to advance algorithm development in a direction that will best cover the needs of the full spectrum of possible users.

Representative Sampling: In order to model affective responses that vary over context and across videos, affective video corpora are needed that include a large number of responses collected from a very large and representative population. However, the number of viewers and their feedback are often limited by our experimental setting and resources. In order to carry out research within the practical constraints of the real world, methods for creating affective video indexing corpora must be both effective—resulting in useful, high-quality corpora—and also efficient with respect to both the time spent and the expense incurred in the development process.

These open issues constitute three dimensions that inform our proposal of guidelines for corpus development for affective video indexing and will steer the development of new corpora to avoid the shortcomings of existing ones. In the next sections, we first introduce the proposed guidelines for affective video corpora and then we discuss how corpora that we have developed have moved progressively towards addressing these limitations.

V. Guidelines for Affective Video Indexing Corpora

In this section, we present a set of corpus development guidelines for affective video indexing. The guidelines are informed by the ground that we have covered thus far, i.e., understanding of emotions and affective response from psychology and techniques available to record it, and also by the general types of multimedia context analysis algorithms that we expect that researchers will be developing with the data sets. We also take into account the limitations of the currently existing corpora, just discussed. From this information, three dimensions emerge that are critical to take into consideration when developing corpora for affective video indexing.

Context of Viewer Emotional Response: Emotional response is complex, and arises not just from the video, but from the context of the video. We consider context to be what the viewer was exposed to before and after the part of the video for which we are interested in the affective impact. Context also includes the people with whom the viewer is watching the video and the viewer's underlying mood and physical state. The complexity cannot be completely controlled, but its impact can be minimized by very explicitly planning the set up in which viewers are exposed to videos. An evaluation protocol should be included that describes exactly what the annotators were asked to do. The protocol ensures that the annotation situation is reproducible should it ever be necessary/desirable to extend the annotations. What is important is to remain firmly focused on how the task is defined so that it is clearly understood that we are trying to predict affective impact on the viewer. Modeling of affect expressed within the video is admitted. However, it should be understood that this is only used as a bridge to infer the ultimate impact on the viewer. It should be clearly stated which parts of the emotional response process, for example, the affective and cognitive components vs. the conative and physiological components. The implications of ignoring the other components should be taken into account.

The formulation of the way in which self-reported emotions are elicited should control for the impact of video before and after the target segment. This includes showing enough of the video. It is important to realize that entertainment video "works" exactly because it takes us as viewers through alternations of mood, or expresses more than one mood at once. Depending on how the video is split up for mood elicitations different (or impartial) viewer responses can be expected. Good handling of context also involves gathering information on the users' underlying mood and physical state.

Personal Variation Among Viewers: Personal variation among viewers has a variety of sources. Some of the personal variation can be dealt with by careful handling of context, as mentioned above. Classical demographic differences are another source of variation. It is important that the target group be defined clearly (e.g., children) so that any existing limitations on user-to-user variability can be as well understood as possible. Narrowing the target group to a very small demographic (e.g., university students in their twenties) should be understood to limit the general applicability of the annotations gathered.

Personal reactions vary according to personal topic preference. It is important to abstract away from topic or the topical interest of viewers: this can be accomplished by using a well balanced data set. Alternately, during data set design, a decision can be made to focus on one particular topic or style of data, which is significant enough to merit study. In any case, the corpus should be as multifunctional as possible: for example, involve enough users so that not only can universal reactions be studied, but also it is possible to study the reactions of different clusters of users that have similar responses.

Effectiveness and Efficiency: In practice, evaluation corpora are always developed under limitations of resources including person power and time. It is important to carefully plan how the corpus development process is handled. Decisions how to most effectively allocate limited resources have a critical effect on the usefulness of the corpus. As much as possible, such decisions should not be made in an arbitrary manner or during the actual process of gathering annotations for the corpus. In order to avoid unnecessarily jeopardizing the usefulness of the corpus, design decisions should be informed by the overall scenario or scenarios for which the data set is being developed. An overall scenario will assure that the type of affective response collected is appropriate for domain. For example, in the case of sports video the expected affective response pattern will be different than

that for television talk shows. Further, the fact that resources are limited means that there is a trade-off between the resolution of the annotations that can be collected and the amount of content that participants can annotate. More sophisticated systems, e.g., PANAS, will lead to very high quality annotations. However, less people are eager to participate in the studies in which the process of response formulation is tedious or otherwise burdensome. Understanding the underlying use scenario, will make it possible to make informed decisions about trade-offs during the corpus design process.

An overall scenario is also helpful, should it become necessary to make further design decisions during the course of corpus development, e.g., decisions how to most effectively use limited resources in the case of unexpected loss of time or budget. Also, if researchers want to reuse the data set later, they have an idea of which uses are appropriate and which uses overstretch what the data set is designed to do. It is also useful to take into account the kind of multimedia content analysis that will be developed and the evaluation measure that will be used. However, it is of critical importance that the corpus be designed to reflect human affective reactions and not be biased to the specific algorithms, or types of algorithms, whose development it is intended to support.

VI. Example Corpora

To demonstrate the application of these guidelines we now turn to discuss concrete examples. Three affective video corpora have been developed using three approaches of increasing sophistication. The lessons learned from each corpus development experience were used to improve the next corpus. As such, the corpora represent a progression that moves towards an ideal corpus for affective video indexing. The annotations for the first dataset were gathered in a laboratory setting. The second dataset was annotated with user affective responses gathered via a Web-based online platform, and the third dataset includes affective responses gathered using an online crowdsourcing platform. The general characteristics of the corpora are presented in Table I for easy reference. The next three sections discuss each in turn.

TABLE 1
The Specifications of the Three Developed Corpora

Laboratory-based responses to movie scenes	
Nr. of participants	10 participants, 3 female, 7 male
Self-reports	free choice words, continuous arousal and valence using sliders using SAM Manikins
Stimuli	Short movie scenes from Hollywood movies
No of videos	64
Selection method	manual
Web-based responses to movie scenes	
Nr. of participants	42 participants, 15 female, 27 male
Self-reports	forced choice from 7 words, arousal and valence on 9 points scale using SAM Manikins
Stimuli	Short movie scenes from Hollywood movies
No of videos	155
Selection method	manual
Crowd-based responses to travelogue videos[1]	
Nr. of participants	32 participants, 11 female, 18 male
Self-reports	forced choice from 11 words, boredom and like-dislike rating on 9 points scale
Stimuli	travelogue videos from "The Big Travel Project"
No of videos	125
Selection method	The whole series

[1] This data set is available to researchers as a part of MediaEval benchmarking data sets. Please, contact MediaEval contacts that can be found at http://www.multimediaeval.org.

A. Movie Scenes Annotated in a Laboratory

Emotional Videos: The first corpus that was developed comprised emotional movie scenes suitable for emotion elicitation and characterization. The affective annotations were gathered via an experiment in which short video clips were shown to participants in a laboratory setting and their physiological responses and emotional self-reports were recorded. Self-reporting included both reporting words, that participants were free to choose themselves, and also reporting continuous arousal and valence using sliders using SAM Manikins.

The physiological responses recorded were peripheral physiological signals generally used for assessing emotions, specifically: Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), which provides heart rate, respiration pattern, and skin temperature. In order to capture facial muscle activity, we also recorded electromyograms (EMG) from the Zygomaticus major and Frontalis muscles. The physiological responses from 8 participants out of 10 were analyzed; signals from two participants were discarded due to technical problems. Subsequently, we calculated audio and visual content features from the videos and studied their correlation with both emotional responses and physiological changes. The GSR, BVP, EMG from Frontalis and Zygomaticus muscles were found to have significant correlation with arousal whereas only facial EMG was among the highly correlated features with valence, i.e., facial expressions were better indicators for valence compared to peripheral physiological signals. EMG from Zygomaticus muscles was found to be correlated with key lighting in the movie scenes and features extracted from BVP were found to be correlated with shot length variation. More detail on the results and analysis of the physiological responses can be found in [63].

Due to the limited time that a participant can spend in each session, a relatively small set of videos, 64 clips from eight movies, were chosen and shown in two sessions. To create this video dataset, we selected video scenes from movies chosen either by using movies selected by similar studies (e.g., [54], [53], [56]), or by choosing recent popular movies. The set of movies included four major genres, namely, drama: The Pianist and Hotel Rwanda; horror: The Ring (Japanese version) and Days Later; action: Kill Bill Vol. I and Saving Private Ryan; comedy: Mr. Bean's Holiday and Love Actually. These four main genres were selected based on a previous study by Wang and Cheong [54]. Although the main genres of the movies were limited to these four, no constraints were imposed on secondary genres, meaning that the movie set also included aspects of, e.g., romance, thriller, sci-fi, history. The scenes that were selected, eight for each movie, had durations of approximately one to two

minutes each and contained an emotional event (as judged by the first author). The complete list of the scenes with editing instructions and descriptions is available online.[1]

Analysis of Assessments: The annotations were collected from ten (three female and seven male) participants ranging in age from 20 to 40 years ($M = 29.3, SD = 5.4$). The difference between arousal and valence scores given by the participants to all the videos was studied by means of a multi-way ANalysis Of VAriance (ANOVA), which was performed on arousal and valence scores considering three factors: the video scenes, the participants, and the order in which the videos were shown to the participants during sessions. The effect of the order in which the videos were presented to the users on the user response was not significant. However, there was a significant difference on average valence scores between different participants ($F(9) = 18.53, p < 1 \times 10^{-5}$) and different videos ($F(63) = 12.17, p < 1 \times 10^{-5}$). There was also a significant difference on average arousal scores between different participants ($F(9) = 19.44, p < 1 \times 10^{-5}$) and different videos ($F(63) = 3.23, p < 1 \times 10^{-5}$). These differences can be attributed to different personal experiences and memories concerning different movies, as well as participants' mood and background.

The development of this corpus provided an important lesson about the personal nature of user-reported affective response and the importance of carefully designing the method for collecting self-reported affective keywords from experiment participants. During the experiments the participants remarked that it was difficult for them to come up with emotion words when watching a video scene. In the end, there was a very low level of consensus among the words that they chose. The overall set of keywords chosen by the participants did not include a high number of instances of basic emotions, e.g., anger, was not very common compared to non-basic emotions, e.g., amusement (see Fig. 3). These observations led to the lesson that giving users complete freedom of choice of response is not particularly helpful in isolating those common aspects of affective response. Instead, it is easier for participants, and yields more stable results, if participants choose from a list of choices. However, the list should not be blindly adopted from the literature, but should be carefully developed for a particular setting using exploratory experiments with participants.

B. Web-Based Annotated Movie Scenes Dataset

Emotional Videos: The development of the second corpus targeted the involvement a larger set of participants. First, a user study was conducted to narrow down the selection of videos to be used as stimuli. This time a more efficient forced-choice self-reporting was used. In order to find videos eliciting emotions from the whole spectrum of possible emotions, a user study was conducted to annotate a set of manually preselected movie scenes. The dataset is drawn from 16 full length Hollywood movies (see [51] for the full list). To create this video dataset, we extracted video scenes from movies selected either according to similar studies (e.g., [54], [53], [56], [63]), or from recent famous movies. A set of 155 short clips, each about one to two minutes long, were manually selected from these movies to form the dataset.

[1]http://cvml.unige.ch/movieList

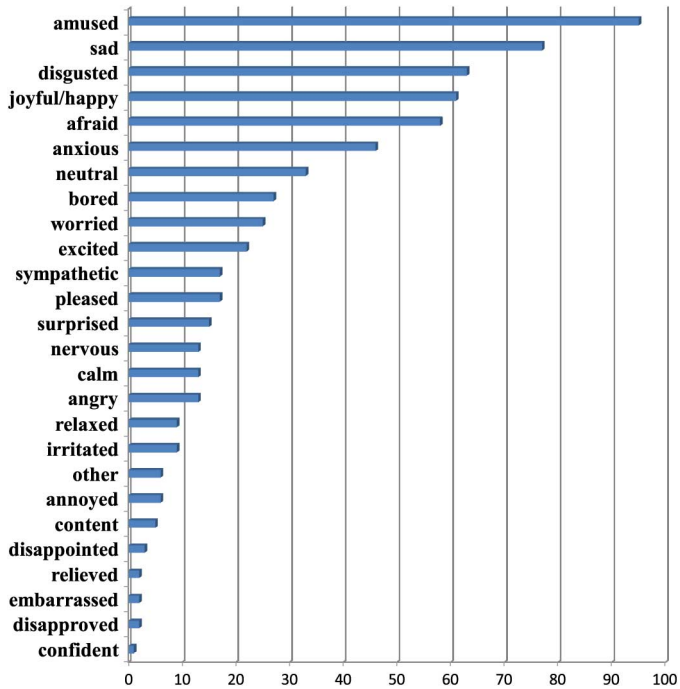

Fig. 3. Total number of keywords reported by 10 participants in response to the 64 video clips of movie scenes in the first corpus.

A Web-based annotation system was developed and deployed in order to collect participants' self-reported affective responses. To use this system, experiment participants sign up and provide personal information including gender, age, and email address. The system also collects information such as cultural background and origin that is used to build a profile of the experiment participants. Providing this information is optional. After watching each video clip, participants express the emotion that they felt using arousal and valence, quantized in nine levels. Participants also choose the emotional label best reflecting the emotions that they feel upon watching the clips. The emotion labels are "afraid", "amused", "anxious", "disgusted", "joyful", "neutral", and "sad". These labels were chosen based on an assessment of the most frequent labels that were used by participants to describe their emotional responses during the development of the first data set (see Section VI-A). Recall that for the first data sets, we asked participants to freely express their emotions, as elicited by movie scenes, with words. The emotional keywords used for the second dataset described in this section are the ones which were used most frequently by participants contributing to the first dataset [63] (see Fig. 3 for the full list of words contributed during the collection of the first dataset).

Analysis of the Self-Reports: Initially, 82 participants signed up to annotate the videos. From these 82 participants, 42 participants annotated at least 10 clips. Participants were from 20 to 50 years old ($M = 26.9, SD = 6.1$). Out of the 42 participants, 27 were male and 15 were female with different cultural backgrounds living in four different continents. We used a

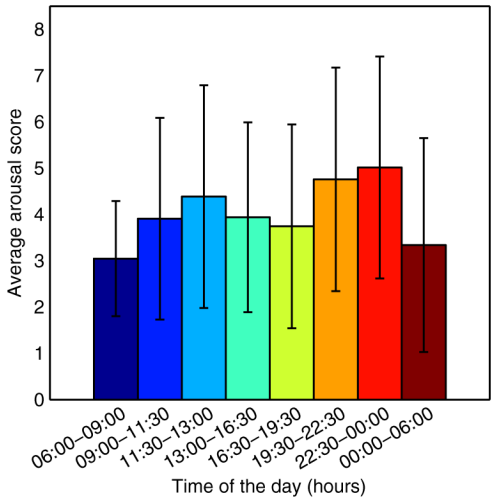

Fig. 4. Average arousal scores in different times of the day. Error bars represent the standard deviation of the ratings.

linear mixed model to test the effect of the time of day on the arousal and valence scores given to the clips. Participant and video clip were considered as random effects and the time of day as fixed effect. The results showed that the average arousal scores are significantly different in different times of the day, i.e., the ANOVA test showed that the coefficients corresponding to different times of the day in the mixed linear model were significantly different from zero ($F(7) = 2.8, p < 0.007$). We did not find the time of day to have significant effect on valence scores. A day was divided into eight time interval, early morning (6:00 to 9:00), morning (9:00 to 11:30), noon (11:30 to 13:00), afternoon (13:00 to 16:30), evening (16:30 to 19:30), late evening (19:30 to 22:30), night (22:30 to 24:00) and after midnight (00:00 to 6:00). The average arousal scores in different time periods are shown in Fig. 4. The average arousal scores given to all videos increases from early in the morning until noon. Then it decreases until it bounces back for late evening and night. The effect of circadian rhythm on self-reported arousal levels reflects the impact of context. Female participants on average gave higher arousal scores to the videos. A Wilcoxon rank sum test showed that the difference between female and male participants' arousal scores was significant ($p = 3 \times 10^{-16}$). These results are in line with the previous findings, e.g., [53], which showed women report stronger emotions than men in response to the same stimuli.

Using the Web-based annotation system had a clear advantage because it increased the number of users from which we were able to collect annotations. We were able to collect enough annotations so that it was meaningful to analyze the variance of the reported arousal scores over times of the day. Also, the Web interface meant that the participants could annotate more video than what was possible in two lab sessions of limited length. Additionally, the development of this corpus led to an important lesson about the context of user annotations. Using the Web interface meant that the environment of the affective response was less controlled. The time of day had a significant impact on the response and it was important to record information about the influence of this factor.

C. Boredom Prediction Dataset

Crowdsourcing for Affective Annotation: In order to reach a broader, more diverse, and larger population, a crowdsourcing platform, Amazon Mechanical Turk (MTurk),[2] was used to gather annotations in the development of the third dataset. The third dataset, initially described in [64], was developed with the aim of supporting research on video processing algorithms capable of predicting viewer boredom. A video dataset has been gathered in the context of the MediaEval[3] 2010 Affect Task for boredom prediction of Internet videos. Using MTurk we rapidly gathered self-reported boredom scores from a large user group that is demographically diverse and also represented our target population (Internet video viewers). Again, the forced choice emotional self-reporting methods were employed. Crowdsourcing practices leverage past work on recruiting subjects and conducting psychological experiments over the Web, which started in the nineties. A set of guidelines were identified by Reips [65] to gather high quality responses from Internet users. Kittur *et al.* [66] showed that when appropriate measures were taken, crowdworkers rating of the quality of Wikipedia articles was comparable with that of experts.

For this work, we adopted a relatively simple, straightforward definition of viewer-experienced boredom. Boredom was taken to be related to the viewer's sense of maintaining focus of attention and is related to the apparent passage of time [67]. Boredom is understood to be a negative feeling associated with viewer perceptions of the viewer-perceived quality (i.e., viewer appeal) of the video being low.

The dataset selected for the corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles.[4] The series consists of 126 videos between two to five minutes in length. This data set was chosen since it represents the sort of multimedia content that has risen to prominence on the Web. Bill's travelogue follows the format of a daily episode related to his activities and as such is comparable to "video journals" that are created by many video bloggers.

Design of Crowdsourcing Task: The third corpus that was developed once again increased the number of annotators and also introduced an even more sophisticated mechanism for context control. The affective responses for this corpus where collected using a large commercial crowdsourcing platform, Amazon Mechanical Turk (http://www.mturk.com). A crowdsourcing platform is an online labor market in which microtasks are offered by requesters and carried out by a pool of human users referred to as "workers". Work in the area of human judgment and decision-making has revealed that there is no difference in the magnitude of the observed effects when experiments are performed using Mechanical Turk and when they are performed with a conventional pool of subjects [68].

The crowdsourcing strategy used for the third corpus was designed based on the existing crowdsourcing literature, for example [66], online articles and blog posts about crowdsourcing

[2]http://www.mturk.com

[3]http://www.multimediaeval.org

[4]http://www.mynameisbill.com

such as "Behind the enemy lines" blog,[5] and also taking into account our past experience regarding collecting annotations in the Web-based experiment. A two-step approach was taken for our data collection. The first step was the pilot that consisted of a single micro-task or Human Intelligent Task (HIT) involving one video. This first HIT was used for the purpose of recruiting and screening MTurk workers as experiment participants. The second step was the main task and involved a series of 125 micro-tasks, one for each of the remaining videos in the collection. Workers were paid 30 US dollar cents for each HIT that they successfully completed.

The pilot HIT contained three components corresponding to responses that were required from the experiment participants that we recruited. The first section contained questions about the personal background (i.e., age, gender, cultural background). The second section contained questions about viewing habits: workers were asked whether they were regular viewers of Internet videos. The third section confirmed their seriousness by asking them to watch the video, select a word that reflected their mood at the moment, and also write a summary. The summary constituted a "verifiable" question, recommended by [66]. The summary offered several possibilities for verification. Its length and whether it contained well-formulated sentences gave us an indication of the level of care that the worker devoted to the HIT. Also, the descriptive content indicated whether the worker had watched the entire video, or merely the beginning. A final question inquired if they were interested in performing further HITs of the same sort. In order not to directly reveal the main goal of the study to workers, the text box for the video summary was placed prominently in the HIT.

The workers were chosen and qualifications were granted for the main task from the participants of the pilot by considering the quality of their description and answers. In the choice of workers, we also strove to maintain a diverse group of respondents. Each HIT in the main study consisted of three parts. In the first part, the workers were asked to specify the time of day. Also the workers were asked to choose a mood word from a drop down list that best expressed their reaction to an imaginary word (i.e., a nonsense word), such as those used in [69]. The mood words were "pleased", "helpless", "energetic", "nervous", "passive", "relaxed", and "aggressive". The answers to these questions gave us an estimate of their underlying mood. In the second part, they were asked to watch the video and give some simple responses to the following questions. They were asked to choose the word that best represented the emotion they felt while watching a video from a second list of emotion words in the drop down list. The emotion list contained the Ekman six basic emotions [55] (namely, "sadness", "joy", "anger", "fear", "surprise", and "disgust") in addition to "boredom", "anxiety", "neutral" and "amusement". For this data set, we had little advance information concerning the emotions that we could expect the video content to trigger in viewers. For this reason, we chose the emotional categories to cover the entire affective space, as defined by the conventional dimensions of valence and arousal [35], supplemented by general information on emotions elicited by film [53]. The emotion and mood word lists contained different items in order to avoid as much as possible that the experiment participants would strongly associate the two. Next, participants were asked to provide a rating specifying how boring they found the video and how much they liked the video, both on a nine-point scale. Finally, they were asked to describe the contents of the video in one sentence.

Analysis of the Ratings: An overall description of the population of the workers who worked on the HIT can be found in [64]. Here we focus on providing details on the ratings. The following questions were asked about each video to assess the level of boredom. First, how boring the video was on nine-point scale from the most to the least boring. Second, how much the user liked the video on the nine-point scale and third how long the video was. Boredom was shown to have on average a strong negative correlation, $\rho = -0.86$ with liking scores. The correlation between the order of watching the videos for each participant and the boredom ratings was also examined. No positive linear correlation was found between the order and boredom score. This means that watching more videos did not increase the level of boredom and, in fact, for two of the participants it lowered their reported boredom level. Additionally, the correlation between the video length and boredom scores was investigated. No positive correlation was found between the boredom scores and videos' duration. We can conclude that longer videos are not necessarily perceived as more boring than the shorter videos.

In order to obtain the dominant mood from the mood words, first the responses of each participant were clustered into the three hours time intervals. In each three hours interval the most frequent chosen mood word was selected as the dominant mood. After calculating the dominant moods, we found that using the implicit mood assessment none of the participants had the "relaxed" as their dominant mood.

The average boredom scores for different dominant moods are shown in Fig. 5. The boredom scores were, on average, lower, i.e., indicated that videos were *more* boring, for viewers in a passive mood and higher, i.e., indicated that videos were *less* boring, in an arguably more active mood such as "energetic", "nervous" and "pleased". Moods were then categorized into positive ("pleased", "energetic", and "relaxed") and negative, ("helpless", "nervous", "passive", and "aggressive") categories. On average, participants gave higher ratings to videos when they were in positive moods. The statistical significance of the difference between ratings in positive and negative moods was examined by a Wilcoxon test and was found significant ($p = 4 \times 10^{-8}$). The effect of the time of day and mood on boredom scores was investigated with a mixed linear model. The fixed effects were mood and time of day, the random effects were video and participant. Unlike the second data set, the effect of the time of day on boredom scores was not significant. This observation can be attributed to the difference between the nature of arousal and perceived boredom, arousal being more correlated with physiological state. Participants' mood had a

[5]http://behind-the-enemy-lines.blogspot.com

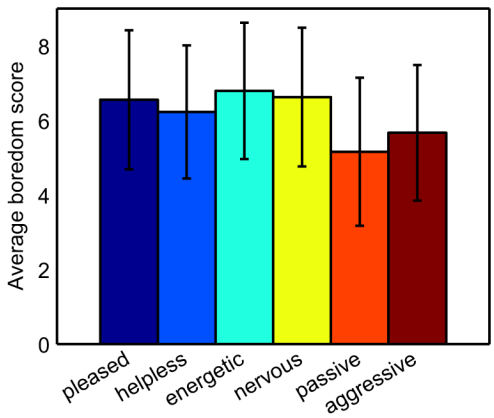

Fig. 5. Average dominant boredom scores reported by viewers experiencing different moods. Error bars represent the standard deviation of the ratings.

significant effect on the ratings; the ANOVA test on the mood coefficients in the mixed linear model showed significant difference from zero $F(6) = 5.75, p < 3 \times 10^{-5}$). The analysis of annotations gathered in this dataset showed the importance of participants' mood which is often not assessed in affective and non-affective assessments.

From the development of this corpus we learned that it is possible to use commercial crowdsourcing to collect a large volume of user affective responses to video. The large number of participants made it possible to analyze sub-groups of participants with particular reactions. Affective response is personal, and varies from individual. However, looking at sub-groups of the population that pattern in the same way could make it possible to isolate the commonalities of viewer response. Variations in the context were addressed by collecting information on the underlying moods of the participants. Even though crowdsourcing is relatively inexpensive, it is still important to plan resources carefully when designing a corpus that uses crowdsourcing to collect affective annotations. A trade-off needs to be made between more annotators, the number of videos annotated and the parts of the HIT design that verify engagement. Advanced planning was necessary to collect the annotations within a set amount of time, since workers may not immediately start working on a HIT once they have qualified. Also, it was necessary to have a large enough pool of workers available to work on the HIT, since some qualified workers do not return to work on the HIT after earning the qualification. In sum, the third corpus addressed all three dimensions of context, personal affective response, and tradeoffs of efficiency and effectiveness.

VII. Discussion and Conclusions

We conclude this paper with some final remarks comparing the three data sets and a brief outlook. To measure inter-annotator agreement, we calculated Krippendorff's alpha (α) for each of the three data sets developed and discussed in this paper: Lab-based $\alpha(arousal) = 0.13$, $\alpha(valence) = 0.49$; Web-based $\alpha(arousal) = 0.22$, $\alpha(valence) = 0.47$, (categorical) $\alpha(emotions) = 0.2$; Crowdsourcing (categorical) $\alpha(emotions) = 0.05$.

Comparison of the second and third data set provides important insights into the distribution of emotion response triggers in video. Recall that the second data set contained manually selected segments of movies that were chosen to create an even distribution in the response patterns that they evoked in viewers. The third data set, on the other hand, does not consist of pre-selected video, rather contains an entire series, i.e., a set of videos as they would occur "in the wild". Both the second and the third data set make use of emotion words that participants selected from a set list (i.e., forced-choice reporting of emotion words). Note that a fixed set of choices can only be used in situations in which it has already been established which emotions can be triggered. The inter-annotator agreement is higher for the manually selected videos than for "in the wild" videos. This difference reflects the fact that "in the wild" videos might not be produced with the intention of evoking strong emotions or that the distribution of the emotion evoked by the video might not be evenly distributed across the categories. In this case of the third data set, the video content is a collection of travelogues, which have a documentary as well as an adventure component to them. They can be expected to evoke emotions, but rather less frequently and also from a more limited set (e.g., "happiness" and "surprise" could be anticipated to be more frequent than "fear"; an emotion such as "anger" might never be triggered.) In sum, comparison of the second and third data sets confirm the importance of understanding the way in which the data collection used for corpus development matches the target data collection to which affective video indexing techniques developed using the data set will ultimately be applied.

This paper has argued that high-quality corpora will help to push forward the state of the art in affective video indexing. In order to realize this predicted potential of affective video corpora, the next step is necessarily the development of additional corpora. If a multitude of corpora can be made available to the research community suitable to support research along the entire spectrum of possible affective video indexing applications, then researchers will have the necessary resources at their disposal to push affective video indexing into the next generation.

Acknowledgment

The authors would like to thank Bill Bowles for granting permission of using his video content for our boredom detection corpus.

References

[1] H.-B. Kang, "Affective content detection using HMMs," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 259–262.

[2] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1114–1122, 2005.

[3] C. H. Chan and G. J. F. Jones, "Affect-based indexing and retrieval of films," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 427–430.

[4] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retriev.*, vol. 4, no. 2, pp. 215–322, 2009.

[5] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proc. ACM Int. Workshop Multimedia Information Retrieval*, 2006, pp. 321–330.

[6] A. F. Smeaton, P. Over, and W. Kraaij, "High-level feature detection from video in TRECVid: A 5-year retrospective of achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin, Germany: Springer-Verlag, 2009, pp. 151–174.

[7] I. Arapakis, J. M. Jose, and P. D. Gray, "Affective feedback: An investigation into the role of emotions in the information seeking process," in *Proc. ACM SIGIR*, 2008, pp. 395–402.

[8] I. Lopatovska and I. Arapakis, "Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction," *Inf. Process. Manage.*, vol. 47, no. 4, pp. 575–592, 2011.

[9] R. W. Picard, Affective computing MIT, Media Laboratory Perceptual Computing Section, Tech. Rep. 321, 1995.

[10] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, 2006.

[11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM TOMCCAP*, vol. 2, pp. 1–19, 2006.

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

[13] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee, "The scholarly impact of trecvid (2003–2009)," *J. Assn. Inf. Sci. Technol.*, vol. 62, pp. 613–627, 2011.

[14] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Int. Conf. Multimedia*, 2000, pp. 105–115.

[15] I. Otsuka, S. Shipman, and A. Divakaran, "A video browsing enabled personal video recorder," in *Multimedia Content Analysis*, ser. Signals and Communication Technology, A. Divakaran, Ed. New York, NY, USA: Springer, 2009, pp. 1–12.

[16] G. J. F. Jones and C. Hau Chan, *Affect-Based Indexing for Multimedia Data*. New York, NY, USA: Wiley, 2012, pp. 321–345.

[17] A. Janin, L. Gottlieb, and G. Friedland, "Joke-o-mat HD: Browsing sitcoms with human derived transcripts," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1591–1594.

[18] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[19] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.

[20] J. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 41–55, 2013.

[21] R. Morris, "The emergence of affective crowdsourcing," in *Proc. ACM CHI '11 Workshop Crowdsourcing and Human Computation*, 2011.

[22] W. Wirth and H. Schramm, "Media and emotions," *Commun. Res. Trends*, vol. 24, no. 3, pp. 3–39, 2005.

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

[24] K. R. Scherer, "Studying the emotion-antecedent appraisal process: An expert system approach," *Cognit. Emotion*, vol. 7, no. 3-4, pp. 325–355, 1993.

[25] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[26] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Netw.*, vol. 18, no. 4, pp. 317–352, 2005.

[27] D. Zillmann, *The Psychology of Suspense in Dramatic Exposition*. New York, NY, USA: Lawrence Erlbaum, 1996, pp. 199–231.

[28] A. I. Nathanson, *Rethinking Empathy*. New York, NY, USA: Lawrence Erlbaum, 2003, ch. 5, pp. 107–130.

[29] D. Zillmann, *Empathy: Affect from Bearing Witness to the Emotions of Others*. New York, NY, USA: Lawrence Erlbaum, 1991, pp. 135–168.

[30] P. Ekman, *Basic Emotions*. New York, NY, USA: Wiley, 2005, pp. 45–60.

[31] J. A. Russell, "Culture and the categorization of emotions," *Psychol. Bull.*, vol. 110, no. 3, pp. 426–450, 1991.

[32] W. Wundt, *Grundzüge der physiologischen Psychologie*. Leipzig, Germany: Engelmann, 1905.

[33] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychol.*, vol. 25, no. 1, pp. 49–59, 1994.

[34] S. Marsella, J. Gratch, and P. Petta, *Computational Models of Emotion*. Oxford, U.K.: Oxford Univ. Press, 2010, ch. 1.2, pp. 21–41.

[35] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.

[36] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.

[37] Y.-H. Yang and H. Chen, "Music emotion ranking," in *Proc. ICASSP 2009*, 2009, pp. 1657–1660.

[38] O. Villon, "Modeling affective evaluation of multimedia contents: User models to associate subjective experience, physiological expression and contents description," Ph.D. dissertation, Université de Nice—Sophia Antipolis, Nice, France, 2007.

[39] A. Kazemzadeh, S. Lee, and S. Narayanan, "Fuzzy logic models for the meaning of emotion words," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 34–49, 2013.

[40] P. G. Hunter, E. G. Schellenberg, and U. Schimmack, "Mixed affective responses to music with conflicting cues," *Cognit. Emotion*, vol. 22, no. 2, pp. 327–352, 2008.

[41] P. G. Hunter, E. Glenn Schellenberg, and A. T. Griffith, "Misery loves company: Mood-congruent emotional responding to music," *Emotion*, vol. 11, no. 5, pp. 1068–1072, 2011.

[42] P. Desmet, *Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products*. Norwell, MA, USA: Kluwer, 2003, ch. 9, pp. 111–123.

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

[44] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect Grid: A single-item scale of pleasure and arousal," *J. Pers. Soc. Psychol.*, vol. 57, no. 3, pp. 493–502, 1989.

[45] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[46] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognit. Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[47] D. Watson and L. A. Clark, "The PANAS-X: Manual for the positive and negative affect schedule–expanded form," 1994.

[48] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder, 2000, "FEELTRACE': An instrument for recording perceived emotion in real time," ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion Sep. 2000.

[49] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA Workshop (ITRW) Speech and Emotion*, 2000, pp. 39–44.

[50] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proc. 2nd ACM Int. Workshop Crowdsourcing for Multimedia*, ser. CrowdMM '13. New York, NY, USA: ACM, 2013, pp. 1–6.

[51] M. Soleymani, J. Davis, and T. Pun, "A collaborative personalized affective video retrieval system," in *Proc. Affective Computing and Intelligent Interaction*, 2009.

[52] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, pp. 42–55, 2012.

[53] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion Elicitation Using Films*, ser. Affective Science. Oxford, U.K.: Oxford Univ. Press, 2007, pp. 9–28.

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

[55] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Pers. Soc. Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.

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

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

[58] R. Plutchik, *A General Psychoevolutionary Theory of Emotion*. New York, NY, USA: Academic, 1980, pp. 3–33.

[59] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 677–680.
[60] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Proc. Affective Computing and Intelligent Interaction*, 2009, pp. 1–7.
[61] R. M. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," in *Multimedia Tools Applications*, 2011, pp. 1–29.
[62] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "A benchmarking campaign for the multimodal detection of violent scenes in movies," in *ECCV Workshops (3)*, ser. LNCS, A. Fusiello, Ed. *et al.* New York, NY, USA: Springer, 2012, vol. 7585, pp. 416–425.
[63] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *Int. J. Semantic Comput.*, vol. 3, no. 2, pp. 235–254, 2009.
[64] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proc. Workshop Crowdsourcing for Search Evaluation, SIGIR 2010*, Geneva, Switzerland.
[65] U.-D. Reips, "The web experimental psychology lab: Five years of data collection on the internet," *Beh. Res. Meth. Instrum. Comput.*, vol. 33, no. 2, pp. 201–211, 2001.
[66] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. ACM SIGCHI Conf. Human factors in Computing Systems (CHI)*, 2008, pp. 453–456.
[67] J. D. Laird*, Feelings: The Perception of Self*, 1st ed. New York, NY, USA: Oxford Univ. Press, 2007.
[68] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgm. Decis. Mak.*, vol. 5, no. 5, pp. 411–419, 2010.
[69] M. Quirin, M. Kazén, and J. Kuhl, "When nonsense sounds happy or helpless: The implicit positive and negative affect test (IPANAT)," *J. Pers. Soc. Psychol.*, vol. 97, no. 3, pp. 500–516, 2009.

Mohammad Soleymani is a Marie Curie Fellow at Imperial College London, where he conducts research on sensor-based and implicit emotional tagging. He received his PhD in computer science from the University of Geneva, Switzerland in 2011. He has worked extensively on assessing emotional reactions in response to video content and developing multimedia techniques to predict these reactions. In the past, he has served as a special session chair, program committee member and reviewer for multiple conferences and workshops including ACM MM, ACM ICMR, and IEEE ICME.

Martha Larson is assistant professor in the Multimedia Information Retrieval Lab at Delft University of Technology. Before coming to Delft, she researched and lectured in the area of audio-visual retrieval at Fraunhofer IAIS and at the University of Amsterdam. Her research interest and expertise lie in the area of speech- and language-based techniques for multimedia information retrieval.

Thierry Pun is head of the Computer Vision and Multimedia Laboratory, full professor at the Computer Science Department, University of Geneva, Switzerland. His current research interests, related to affective computing and multimodal interaction, concern physiological and behavioral signals analysis for affective state assessment, affective gaming and learning, affect in social media, brain-computer interaction, multimodal interfaces for blind users and for the elderly. He has authored or co-authored over 300 full papers as well as eight patents. He is in the editorial boards of the International Journal on Image & Video Processing, and Advances in Multimedia. He was one of the general chairs of ACII - Affective Computing and Intelligent Interaction 2013 in Geneva.

Alan Hanjalic is a professor and head of the Multimedia Signal Processing Group at the Delft University of Technology, The Netherlands. His research focus is on multimedia information retrieval, recommender systems and social media analytics. Prof. Hanjalic is an elected member of the IEEE Technical Committee on Multimedia Signal Processing and an appointed member of the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA. He has been a member of Editorial Boards of five scientific journals in the multimedia field, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the International Journal of Multimedia Information Retrieval. He has also held key positions in the organizing committees of leading multimedia conferences, including the ACM Multimedia, ACM ICMR and IEEE ICME.