Affective Video Content Representation and Modeling

Alan Hanjalic, *Member, IEEE,* and Li-Qun Xu, *Member, IEEE*

***Abstract*—This paper looks into a new direction in video content analysis – the representation and modeling of *affective video content*. The affective content of a given video clip can be defined as the intensity and type of feeling or emotion (both are referred to as *affect*) that are expected to arise in the user while watching that clip. The availability of methodologies for automatically extracting this type of video content will extend the current scope of possibilities for video indexing and retrieval. For instance, we will be able to search for the funniest or the most thrilling parts of a movie, or the most exciting events of a sport program. Furthermore, as the user may want to select a movie not only based on its genre, cast, director and story content, but also on its prevailing mood, the affective content analysis is also likely to contribute to enhancing the quality of personalizing the video delivery to the user. We propose in this paper a computational framework for affective video content representation and modeling. This framework is based on the *dimensional approach to affect* that is known from the field of psychophysiology. According to this approach, the affective video content can be represented as a set of points in the two-dimensional *(2-D) emotion space* that is characterized by the dimensions of *arousal* (intensity of affect) and *valence* (type of affect). We map the affective video content onto the 2-D emotion space by using the models that link the arousal and valence dimensions to low-level features extracted from video data. This results in the arousal and valence time curves that, either considered separately or combined into the so-called *affect curve*, are introduced as reliable representations of expected transitions from one feeling to another along a video, as perceived by a viewer.**

***Index Terms*—Affective video content analysis, video abstraction, video content modeling, video content representation, video highlights extraction.**

I. Introduction

DIGITAL VIDEO collections are growing rapidly in both the professional and consumer environment, and are characterized by a steadily increasing capacity and content variety. Since searching manually through these collections is tedious and time-consuming, transferring the search and retrieval tasks to automated systems becomes crucial for being able to efficiently handle stored video volumes. The development of such systems is based on the *algorithms for video content analysis*. These algorithms are built around the models bridging the gap between the syntax of the digital video data stream (captured in the so-called *low-level features*) and the semantic meaning of that stream. Using the information that is extracted from a video by these algorithms, digital video data can be indexed, classified, filtered or organized automatically based on semantic criteria.

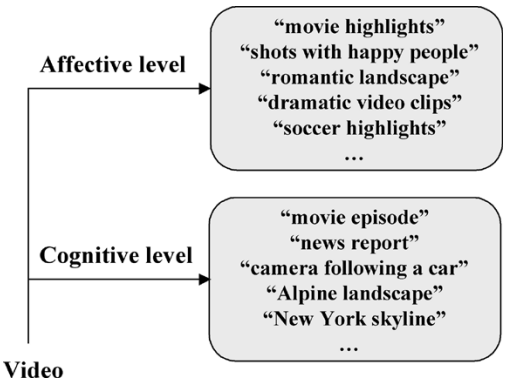

Fig. 1. Overview of two different levels of video content perception, analysis, and retrieval.

The semantic meaning of a given video clip is not unique, as the content of this clip can be perceived in many different ways. Clearly, each way of perceiving video content requires a particular type of information in order to index, classify, filter or organize the video collection correspondingly. As depicted in Fig. 1, we differentiate between two basic levels of video content perception, hence two different levels of analyzing and retrieving video content.

- Cognitive level.
- Affective level.

An algorithm analyzing a video at cognitive level aims at extracting information that describes the "facts," e.g., the structure of the story, the composition of a scene, and the objects and people captured by the camera. For example, these facts can include the labels such as "a panorama of San Francisco," an "outdoor" or "indoor" scene, a broadcast news report on "Topic X," a "dialog between person A and person B," or the "fast breaks," "steals," and "scores" of a basketball match. Most of the worldwide research efforts in the field of video content analysis have been invested so far in raising the efficiency and reliability of analyzing the video content at cognitive level. Good overviews of, and references to the results of these efforts can be found in [7], [13], and [21].

Little research effort has been invested so far in extracting the information that describes the *affective content* of a video. This content can be defined as the amount and type of *affect* (feeling or emotion) that are contained in video and expected to arise in

Manuscript received August 31, 2001; revised June 18, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sankar Basu.

A. Hanjalic is with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: A.Hanjalic@ewi.tudelft.nl).

L.-Q. Xu is with the Broadband Applications Research Centre, BT Research Venturing, Martlesham Heath, Ipswich IP5 3RE, U.K. (e-mail: li-qun.xu@bt.com).

Digital Object Identifier 10.1109/TMM.2004.840618

users while watching that video. This *expected* feeling or emotion can be seen as the one that is either intended to be communicated toward the audience (from video program directors), or that is likely to be elicited from majority of the audience who are watching the particular video clip. To illustrate the former, we use the quote of I. Maitland [25], the Emmy-Award-winning director and editor: "*It is the filmmaker's job to create moods in such a realistic manner that the audience will experience those same emotions enacted on the screen, and thus feel part of the experience.*" The expected affective response of a broad audience can best be illustrated by the example of a sport broadcast: A score (goal) in a soccer match can generally be considered a highly exciting event, just like the finish of a swimming competition or the sprint over the last 50 m in a running contest.

At this stage it is worthwhile emphasizing that the affective content of a video does not necessarily correspond to the affective response of a particular user to this content. In other words, the *expected* feeling or emotion as described above should not be mixed up with the *actual* feeling or emotion that is evoked in a user while watching video. The expected affective response can be considered objective, as it results from the actions of the movie director, or reflects the more-or-less unanimous response of a general audience to a given stimulus. Opposed to this, the perceived feeling or emotion is highly subjective and context-dependent. Therefore, it may be very different from the expected one and may also vary from one individual to another. For instance, the same soccer television broadcast may make the winning team's fans happy, the losing fans sad, and elicit no emotions at all from an audience that is not interested in soccer. The relation between the expected and the subjective affective responses (e.g., marking a horror movie with the label "funny" for those people who always laugh while watching such movies) and the information about the context (e.g., winning or losing soccer fan) can be taken into account, for instance, by generating the profile of a particular user. This profile can then be used to map the expected affective response to a given stimulus onto the user-specific affective response to that stimulus.

We propose in this paper a computational framework for affective video content representation and modeling. The representation part of the framework consists of a set of curves that reliably depict the expected transitions from one feeling to another along a video, as elicited from a general user. The modeling part addresses the problem of computing the values of the content representation curves on the basis of low-level features extracted from video.

This paper is organized as follows. In Section II, we discuss the importance of extending the research in the field of video content analysis from the cognitive to the affective level, which allows for a number of new or enhanced video indexing and retrieval applications. In Section III we elaborate on the *dimensional approach to affect* that is known from psychophysiology and that provides the fundamentals of the proposed framework. The detailed framework is then presented in Section IV (representation part) and Section V (modeling part), together with the validation using real video program data. Conclusions and recommendations for future research are given in Section VI.

II. Why Affective Video Content Analysis?

A. Personalized Video Delivery

In view of the rapidly growing technological awareness of the average user, the availability of automated systems that can optimally prepare data for an easy access by the user becomes crucial for the commercial success of consumer-oriented multimedia databases. The minimum expected capabilities of such systems will definitely evolve beyond the pure automation of retrieval processes: an average user will require more and more from his electronic infrastructure at home. In the particular case of video storage systems, this "more" can directly be interpreted as *personalized video delivery*. Since video storage system at home will soon become a necessary buffer for the hundreds of television channels reaching one's home, the system module handling the stored video data will increasingly be expected to take into account the preferences of the user and to filter and organize the stored content accordingly. The systems currently available for personalized video delivery usually filter the programs on the basis of information like, in the case of a movie, the genre, cast, director and story (script) content. As the user preferences in this case are also largely determined by the prevailing mood of a movie, then any information regarding this mood (obtainable by analyzing the types and intensities of feelings or emotions along a video) is likely to improve the quality of personalized video delivery.

B. Video Indexing Using Affective Labels

The availability of methods for automatically extracting the affective video content will extend the current scope of possibilities for video indexing and retrieval. The evidence reported by Picard [25] is that finding photographs having a particular mood was the most frequent request of advertising customers in a study of image retrieval made with Kodak Picture Exchange [27]. One can easily extend this result to video collections as well: an average user will often search for the "funniest," "most sentimental," or "most thrilling" fragments of a movie, as well as for the "most exciting" segments of a sport event.

C. Video Highlighting

Although the highlights generally stand for the most interesting parts of a video, the definition of what is "interesting" may vary widely across video genres and for different applications. For instance, while a highlight of a news program may be determined by the novelty and impact of the news (e.g., "breaking news," "headline news"), the criteria for highlight extraction from a home video are rather content-dependent, like "where my baby walked for the first time." The ability to analyze video at affective level will broaden the possibilities for highlights extraction in a number of application contexts, such as automated movie trailer generation and sport broadcast summarization.

A movie trailer is a concatenation of movie excerpts that last only for several tens of seconds but are capable of commanding the attention of a large number of potential cinema goers and video on-demand users. Analyzing a movie at affective level can provide valuable clues about which parts of the movie are

most suitable for being an element of the trailer. This is because emotion plays a primary role when processing mediated stimuli [9]. The emotion (affective content) influences the attention of a user and his or her evaluation and memory for the facts (cognitive content). Consequently, the perception of the affective content interferes with the perception of the cognitive content and influences a user's reactions to the cognitive content, such as liking or not-liking, enjoyment and memory. Since memory is the most important factor when creating a trailer, it is worthy to notice that memory for highly emotional and, in particular, highly arousing video fragments has been proven to last longer than the memory for less-emotional video clips [16], [17]. If the information on the affective video content is available, creation of movie trailers can be performed fully automatically. Also, the trailers can be generated remotely at a user's home, for each movie downloaded by the home digital video storage system.

Previous approaches to automated highlights extraction in sport video were usually based on the development of domain-specific models for predefined events (e.g., goals in soccer, home runs in baseball, fast breaks and steals in basketball, etc.) that are supposed to be interpreted by the users as highlights [3], [15], [22]. The need for event modeling not only makes the highlight extraction technically and semantically a complex task in many broadcasts, but it also requires the development of a separate highlights-detecting algorithm for each particular sport program genre. Since it is realistic to assume that each highlight event (e.g., goal, touchdown, home run, the finals of a swimming competition, and the last 50 meters in a running contest) induces a steady increase in a user's excitement, an alternative to the domain-specific approach could be to search for highlights in those video segments that excite the users most. In this way, generic methods for highlights extraction could be developed that are independent of the type of events appearing in a particular sports program genre and the differences in event realization and coverage.

III. Dimensional Approach to Affect

As studied by Bradley [5], Lang *et al.* [19], Osgood *et al.* [24], Russel and Mehrabian [28], affect has three basic underlying dimensions.

- Valence (V).
- Arousal (A).
- Control (Dominance) (C).

Valence is typically characterized as a continuous range of affective responses or states extending from pleasant or "positive" to unpleasant or "negative" [8], while arousal is characterized by affective states ranging on a continuous scale from energized, excited and alert to calm, drowsy or peaceful. We can also say that arousal stands for the "intensity" of emotion, while valence can be related to the "type" of emotion. The third dimension – control (dominance) – is particularly useful in distinguishing among emotional states having similar arousal and valence (e.g., differentiating between "grief" and "rage") and typically ranges from "no control" to "full control". Consequently, the entire scope of human emotions can be represented as a set of points in the three-dimensional (3-D) VAC coordinate space.

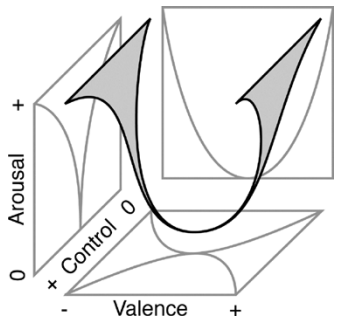

Fig. 2. Illustration of the 3-D emotion space (from Dietz and Lang [9]).

While we could tend to assume that the points corresponding to different affective states are equally likely to be found anywhere in the three-dimensional VAC coordinate space, psychophysiological experiments show that only certain areas of this space are actually relevant. These experiments typically include measurements of affective responses of a large group of subjects to calibrated audio-visual stimuli collected in the *International Affective Picture System* (IAPS, Lang *et al.* [20]) and the *International Affective Digitized Sounds* system (IADS, Bradley and Lang [6]). Subjects' affective responses to these stimuli can be quantified either by evaluating their own reports, e.g., by using the Self-Assessment Manikin ([18]) or by measuring physiological functions that are considered related to particular affect dimensions. For example, heart rate reliably indexes valence, while skin conductance is associated with arousal. It was found that the heart rate accelerates as a reaction to pleasant stimuli, while unpleasant stimuli cause the heart rate to slow down [8], [10], [12]. Also, an increase in arousal causes the sweat glands to become active and the skin conductance responses larger and more frequent [8], [14]. While IAPS and IADS are specially created to evoke a wide range of emotions with their audio-visual content, the three-dimensional surface circumventing the affective responses after their mapping onto the corresponding points in the 3-D VAC coordinate system is roughly parabolic. An idea about the shape of the surface can be obtained from the illustration in Fig. 2. The parabolic shape becomes logical if we realize that there are relatively few or even no stimuli that would cause an emotional state characterized by, for instance, high arousal and neutral valence, or high valence accompanied by low arousal [9].

The dimensional approach to representing emotion as described above can play an important role in the development of "affective" agents that serve as mediators between the computer and user, and involve the user in an interaction with the computer in the same way as he/she interacts with other humans. Since human-to-human interaction is strongly determined by emotions, an affective agent is able to sense, synthesize, and express emotions. For example, Dietz and Lang [9] use the parabolic surface from Fig. 2 as the basis for assigning a temperament, mood and emotion to an affective agent, thus defining the "personality" of that agent. The temperament is a fixed point in the space that defines the "at rest" state of the agent (its rudimentary personality). While the temperament is static, the points corresponding to the mood and emotion of the agent can move freely within the space. The position of the emotion point gives

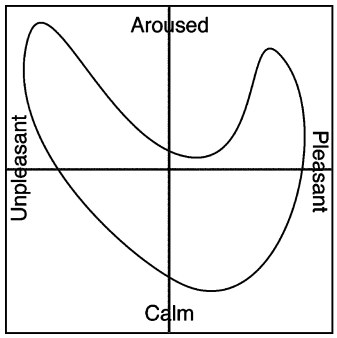

Fig. 3. Illustration of the 2-D emotion space (from Dietz and Lang [9]).

rise to the expressions of the agent and determines its current affective state. Further, the emotion point gravitates toward the position of the mood that, again, moves through the space relatively slowly, is mainly pulled by emotional events and gravitates toward the position of the temperament. The dynamics of the system is therefore influenced by both the agent's current affective state and its temperament.

IV. Affective Video Content Representation

A. Two-Dimensional (2-D) Emotion Space

As can be seen from Fig. 2, the effect of the control dimension becomes visible only at points with distinctly high absolute valence values. This effect is also quite small, mainly due to a rather narrow range of values belonging to this dimension. Consequently, it can be said that the control dimension plays only a limited role in characterizing various emotional states. As a matter of fact, Greenwald *et al.* [12] have shown that valence and arousal account for most of the independent variance in emotional responses. This is especially true for the problem to be addressed in this paper – the extraction of the affective content from a video. Numerous studies of human emotional responses to media have shown that "emotion elicited by pictures, television, radio, computers, and sounds can be mapped onto an emotion space created by the arousal and valence axes" [9]. For this reason, we neglect the control dimension and consider the arousal and valence dimensions only. Instead of the three-dimensional surface introduced in the previous section, the relevant emotion space for the purpose of affective video content analysis is reduced to the projection of this surface onto the arousal-valence plane. Fig. 3 shows an illustration of the resulting 2-D emotion space. The parabolic contour is generated to circumvent the scatter plot of affective responses with respect to arousal and valence only, which were collected using the IAPS and IADS stimuli. It is expected that the affective states extracted from a video can be represented as the points within this contour.

B. Arousal, Valence, and Affect Curve

By computing the arousal and valence values along a video the arousal and valence time curves can be obtained. We introduce these curves, either considered separately or combined into the so-called *affect curve*, as suitable representations of the affective content of a video in view of the applications described in Section II.

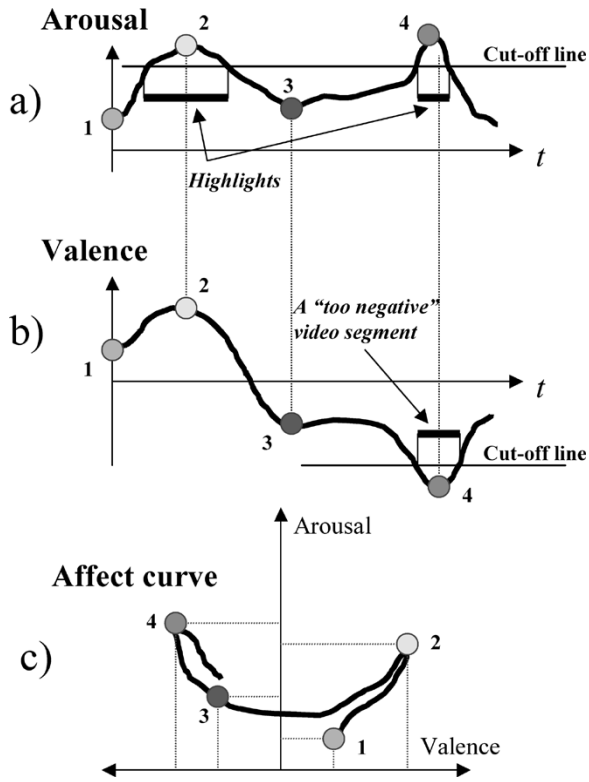

Fig. 4. Illustration of the arousal, valence, and affect curve.

The *arousal time curve* indicates how the intensity of the emotional load changes along a video, and depicts the expected changes in user's excitement while watching that video. In this sense, the arousal curve is particularly suitable for locating the "exciting" video segments. On the basis of the arousal time curve we can generate a video abstract containing the highlights in a desired length. Namely, given the maximum abstract length N in frames, a horizontal line can be drawn cutting off the peaks of the curve in such a way that the number of frames covered by the peaks is not larger than N. This is illustrated in Fig. 4(a).

The *valence time curve* depicts the state changes in the type of feelings or emotions contained in a video over the time. As such, this curve mimics the expected changes of "moods" of the user while watching a video. Using the valence time curve we can also determine the "positive" and "negative" video segments with respect to the expected type of feeling that is evoked in the user during these segments. This information can serve to match the video to personal preferences of the user, but also to automatically perform "censorship" tasks, that is, to remove all segments from a video that are "too negative" for certain groups of the audience. As illustrated in Fig. 4(b), such segments may be searched among those for which the valence curve reaches local minima. The arousal and valence time curves can be combined into the *affect curve*. This curve is composed of the value pairs of the arousal and valence time curves that are taken per time stamp of the video and mapped onto the corresponding points of the 2-D emotion space [Fig. 4(c)].

The affect curve can be seen as the most complete representation of the affective content of a video, which can be obtained automatically. This curve can be interpreted in various

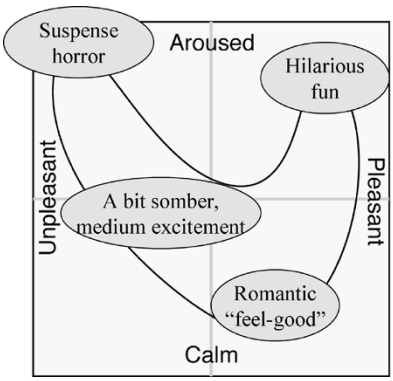

Fig. 5. Illustration of the possibility for video content indexing and retrieval at affective level.

ways and used for numerous applications related to video content representation and retrieval at the affective level. For instance, assuming that the affect curve has already been computed for a given video, an arbitrary temporal segment of that video can automatically be indexed with respect to the affective states through which the corresponding part of the affect curve passes. Indexes can be provided in the form of labels that are assigned a priori to different regions of the 2-D emotion space, as illustrated in Fig. 5. Also, the area of the 2-D emotion space in which the curve traverses most of the time corresponds to the dominant affective state ("prevailing mood") of a video. This can be highly useful for automatically classifying a video into different affective genres. Further, the affect curve may directly serve as a criterion for filtering the incoming videos according to a user's preferences. Namely, an affect curve representing a user's preferences can be obtained by simply combining the affect curves of all programs that the user has selected in the past (in the learning phase of the system). Filtering an incoming video according to this user's preferences is then nothing more than matching the affect curve of the incoming video with the affect curve describing the user's preferences.

V. Affective Video Content Modeling

In order to obtain the affective content representation as described in the previous section, models need to be developed for the arousal and valence time curve. These models fulfill the tasks of deriving arousal and valence values from the values of low-level features computed in a video. In Section V-A, we introduce the basic criteria that need to be taken into account during the model development. Then, in Section V-B, we elaborate on the possibilities for establishing relations between the affect dimensions and low-level features. Finally, we propose models for the arousal and valence time curve and experiment with these models on a number of video excerpts from movies and soccer television broadcasts.

A. Criteria for Developing Affect Models

As arousal and valence are psychological categories, their models need to be psychologically justifiable. To achieve this, we introduce the following three criteria that a model for the arousal, valence or affect curve should satisfy.

- Comparability.
- Compatibility.
- Smoothness.

The first criterion (*Comparability*) ensures that the values of the arousal, valence and the resulting affect curve obtained in different videos for similar types of events are comparable. This criterion obviously imposes normalization and scaling requirements when computing the time curves. The second criterion (*Compatibility*) ensures that the affect curve covers an area in the valence-arousal coordinate system, the shape of which roughly corresponds to the parabolic-like contour of the 2-D emotion space. The third criterion (*Smoothness*) accounts for the degree of memory retention of preceding frames and shots [1]. It ensures that the perception of the content, and consequently the mediated affective state, does not change abruptly from one video frame to another but is a function of a number of consecutive frames (shots).

B. Feature Selection

Little is known regarding the relations between the low-level features and affect. While the problem of bridging the semantic gap remains very hard in the case of cognitive video content analysis, the magnitude of this problem in the affective case is even bigger. The reason for this is that in the cognitive case the low-level features describe aspects of a real entity, e.g., the choice of the color *red* as one of the features to characterize a red car. In the affective case, however, we need to relate the low-level features to something rather abstract, such as feeling or emotion. In the context of this paper, we are particularly interested in the relations between low-level features and the affect dimensions of arousal and valence.

One of the most extensively investigated visual features in the context of affective video content analysis is motion. Research results show that motion in a television picture has a significant impact on individual affective responses. This has been realized also by film theorists who contend that motion is highly expressive and is able "to evoke strong emotional responses in viewers" ([2], [11]). In particular, Detenber *et al.* [8] and Simmons *et al.* [29] investigated the influence of camera and object motion on emotional responses of humans and concluded that an increase of motion intensity on the screen causes an increase in arousal. The type of emotion (represented by the sign of valence) was found independent of motion: if the mood of a test person was "positive" or "negative" while watching a still picture, the "sign" of the mood will not change if a motion is introduced within that picture.

Based on the results obtained by Murray and Arnott [23], as well as those reported by Picard in [25] and [26], various vocal effects present in the sound track of a video may bear broad relations to the affective content of that video. In terms of affect dimensions, the loudness (signal energy) and speech rate (e.g., faster for fear or joy and slower for disgust or romance) are often being related to the arousal, while the inflection, rhythm, duration of the last syllable of a sentence, voice quality (e.g., breathy or resonant), as well as the pitch-related features (pitch average, pitch range and pitch changes), are commonly related to valence [23], [25].

Finally, we found the varying shot length to be a good example of an editing effect that can be put in relation to the affective content of video. Namely, the patterning of shot lengths [1] is a popular tool for the director to create the desired pace of action (e.g., in a movie). The director typically chooses for shorter shot lengths in movie segments that are to be perceived by the viewers as those with a high tempo of action development, or to create stressed, accented moments. As opposed to this, longer shots are typically used to de-accentuate an action [4]. In this sense, the varying shot lengths can be linked to the intended changes in the magnitude of arousal that is evoked in the audience along a movie. Note that regarding the pace at which the video content is offered to a viewer, an increase in shot-change rate is likely to have a similar impact on a viewer's arousal as an increase in the overall motion activity.

Wide variations in shots lengths can also be a good indication of how the director of a live broadcast responds to interesting events. We can explain this on the example of a soccer match that is broadcasted most of the time using one camera that covers the entire field. The director switches from one to another camera (e.g., by zooming onto a particular event, the bench or the spectators) only occasionally, which results in rather long shots. However, whenever there is a goal, or an important break (e.g., due to foul play, free kick, etc.), the director immediately increases the rate of shot changes trying to show everything that is happening on the field and among the spectators at that moment. In this way, any increase in shot-change rate during a live broadcast is likely to be related to the director's response to an increase in the general arousal evoked in the sport arena.

C. Model for Arousal

We start our approach to arousal modeling by considering the function $G_i(k)$ that models the changes in the arousal over the frames k as revealed by the feature i. This function can be interpreted as one of the components of the arousal time curve. Namely, it has been realistically expected that no single feature can reveal the complete variations of arousal along a video. For instance, an increase in arousal during a soccer television broadcast is detectable at some places through the cheering crowd (changes in sound energy) and at some other places through an increase in shot-change rate (e.g., a break due to a foul play). Therefore, we model the arousal time curve $A(k)$ in general as a function of N components $G_i(k)$

$$A(k) = F(G_i(k), i = 1, \ldots, N). \tag{1}$$

Here, the function F serves to integrate the contributions of all the components $G_i(k)$ in the overall course of arousal along a video. In order for the function F to satisfy the criteria of comparability and smoothness, these criteria need to be satisfied first by each component time function $G_i(k)$. This requirement can also be justified by the fact that each function $G_i(k)$ is an (elementary) arousal function by itself.

We now search for the appropriate form of the function (1) and investigate its ability to reliably represent the arousal changes along a video. For this purpose, we use three sample low-level features that were selected on the basis of the discussion in Section V-B.

a) *The motion component*, obtained on the basis of the overall motion activity measured between consecutive video frames.
b) *The rhythm component*, obtained by investigating the changes in shot lengths along the video.
c) *The sound energy component*, obtained in synchronization with video frame interval by computing the total energy in the sound track of a video.

The above features were selected to represent the arousal stimuli contained in different modalities of video (visual and audio) and those revealing the influence of video authoring (editing). In this sense, we expect the contributions to the course of arousal originating from different features to be largely independent of each other. In the following, we first model the component time curves $G_i(k)$ for the selected features. Then, we propose a function F integrating these three components, and evaluate it on a number of representative test sequences.

1) Motion Component: We start the computation of the motion component of the arousal function (1) by computing the motion activity $m(k)$ at each video frame k. Motion vectors are computed using the standard block-based motion estimation [13] between adjacent two frames k and $k+1$. The motion activity value is then found as the average magnitude of all (B in total) motion vectors $\overrightarrow{v}_i(k)$, normalized by the maximum possible length of a motion vector $|\overrightarrow{v}_{\max}|$

$$m(k) = \frac{100}{B|\vec{v}_{\max}|}\left(\sum_{i=1}^{B}|\vec{v}_i(k)|\right)\%. \tag{2}$$

Note that the motion activity values (2) are scaled to the range between 0% and 100%, a range that will be imposed also for other model components so they can be combined with each other on the same basis, but also for the resulting arousal levels to be expressed in percentages. In this way, we create a solid basis for the fulfillment of the *Comparability* criterion.

In view of the *Smoothness* criterion, the obtained motion activity time curve is not directly suitable for being a component of the arousal model. First, the value (2) may quickly fluctuate within the same shot. Second, motion-activity values may fluctuate in different ranges for two consecutive shots (e.g., total motion activity within a close-up shot is much larger than that in a shot taken from a large distance) which results in "jumps" of these values from one range to another at shot boundaries. Third, measuring motion activity for the consecutive frames will encounter unavoidably the high peaks or other noises at shot boundaries and locations of other editing effects as well. In order to fulfill the *Smoothness* criterion the $m(k)$ is convolved with a sufficiently long smoothing window. We use the Kaiser window $K(l_1, \beta_1)$ of the length l_1 and the shape parameter β_1 for this purpose. This window is illustrated in Fig. 6.

We demonstrate the effect of the smoothing operation in Fig. 7, where a video segment consisting of three consecutive shots of a typical soccer match is considered. The two shot boundaries can be easily recognized as the sharp peaks around frames 200 and 300 of the motion activity function $m(k)$ in Fig. 7(a). The first and second shot are characterized by a high motion activity, corresponding to close-up shots of players running on the field. The third shot was taken by a camera

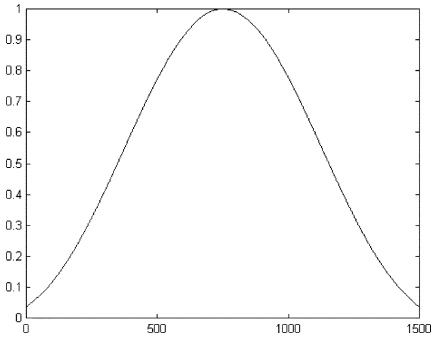

Fig. 6. Example of the Kaiser window $K(l, \beta)$ of the length $l = 1500$ and shape parameter $\beta = 5$.

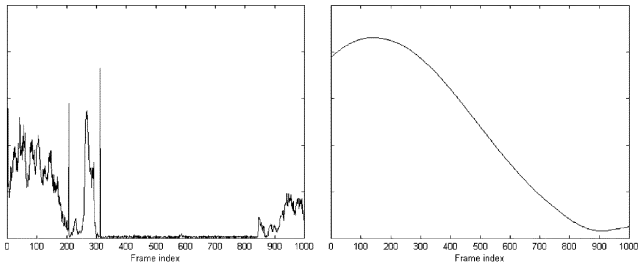

Fig. 7. (a) Example of the motion activity curve $m(k)$. (b) Motion activity curve resulting from a convolution with a Kaiser window.

mounted on a high ground with a wide view of the field, hence the overall motion activity obtained is rather low initially. The changes toward the end of the third shot take place when the camera is maneuvered to view the previously covered part of the field in the course of the game. The first two shots belong to an exciting segment of a soccer broadcast (goal chance). Starting from the second shot change, the game becomes stable and the level of excitement decreases. However, the increase and decrease of a user's excitement cannot change abruptly. While a user's excitement will reach its peak somewhere during the series of close-up shots, it will start to descend gradually, as the game becomes stable. Gradual reduction in the level of excitement will continue after the second shot change since the user needs time to recover from previous exciting events (inertia). At last, when the course of the game becomes more dynamical around frame 850, the excitement of the user will start to rise, though with a certain delay – again due to the inertia of human affective states. As can be seen from Fig. 7(b), the smoothed motion activity curve is much more likely to mimic the variations in a user's excitement, as described above.

We adopt the smoothed motion activity curve as the motion component $G_1(k)$ of the arousal time curve (1). We represent this component analytically as

$$G_1(k) = \frac{\max\limits_k(m(k))}{\max\limits_k(\widetilde{m}(k))}\widetilde{m}(k)\%. \tag{3}$$

Here, $\widetilde{m}(k)$ is the result of the convolution of the curve $m(k)$ with a smoothing window, that is $\widetilde{m}(k) = m(k) * K(l_1, \beta_1)$. Scaling the curve $\widetilde{m}(k)$, as indicated in (3), serves to put the values $G_1(k)$ back inside the original value range (0%–100%).

2) Rhythm Component: Similar to the analysis of the motion activity in the following we aim at obtaining a curve that is a function of the frame index k and that reveals a connection between a viewer's arousal and the time-varying shot lengths. We start modeling the influence of the shot-change rate on a viewer's arousal by defining the function $c(k)$

$$c(k) = 100e^{((1-(n(k)-p(k)))/\delta)}\%. \tag{4}$$

Here, $p(k)$ and $n(k)$ are the positions (frame indexes) of the two closest shot boundaries to the left and right of the frame k, respectively, and the parameter δ is the constant determining the way the $c(k)$ values are distributed on the scale between 0% and 100%. As illustrated on the example in Fig. 8(a), the curve $c(k)$ is typically a step curve, with each step corresponding to video segment between two shot boundaries and with the height of each step being inversely related to the interval between the boundaries: the shorter the interval, the higher the value $c(k)$. Again, due to incompatibility of vertical edges in $c(k)$ with the *Smoothness* criterion, we convolve the $c(k)$ curve with the same smoothing window as in the case of motion activity. Scaling the convolution result back to the original value range results in the function that we adopt as the rhythm component $G_2(k)$ of our arousal model (1), which is illustrated in Fig. 8(b)

$$G_2(k) = \frac{\max\limits_k(c(k))}{\max\limits_k(\widetilde{c}(k))}\widetilde{c}(k)\%, \quad \text{where} \quad \widetilde{c}(k) = K(l_1, \beta_1) * c(k). \tag{5}$$

3) Sound Energy Component: As the third component of the proposed arousal model, the sound energy contained in the audio track of a program is considered. One energy value is computed for the time length of each video frame. Thus the number s of audio samples used to compute this value is determined as the ratio between the audio sampling frequency (typically 44.1 kHz for CD quality) and the video frame rate. The power spectrum is computed for each consecutive segment of the audio signal containing s samples. An equivalent of the sound energy value $e(k)$ is then computed by adding up all spectral values.

We again apply the same Kaiser window as in previous sections to smooth out the originally "rough" time curve $e(k)$. However, unlike the other two arousal components, sound energy is dependent on the volume level at which the audio track is recorded. Since neglecting this fact would result in sound energy time curves that are not comparable over different videos, we proceed as follows. First, we scale the energy time curve obtained after convolution to the range between 0 and 1. Then, we weight the obtained curve according to its mean value. If the curve is characterized by only a few highly distinguishable peaks, then its mean value is lower than in the case where the curve homogeneously covers the entire value range. Since, in the first case, it is likely that video contains several highly exciting events, these peaks should play a significant role in shaping the final arousal time curve. In the second case, however, the presence of exciting events is uncertain. Then, due to ambiguity related to the recording volume level, the influence of the energy component on shaping the arousal time curve is kept limited. With this in mind, with $\widetilde{e}(k) = K(l_1, \beta_1) * e(k)$

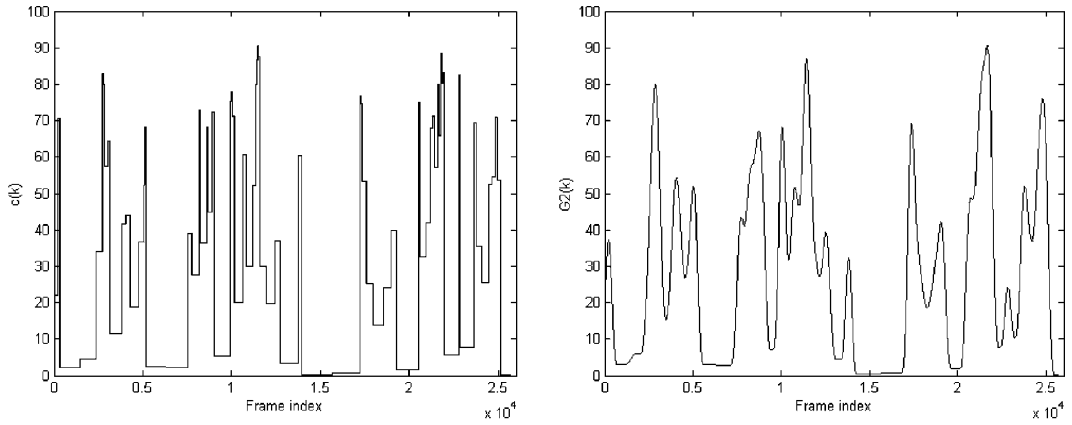

Fig. 8. (a) Example of the curve $c(k)$. (b) Corresponding curve $G_2(k)$.

and with W being the length of the analyzed video in frames, we define the sound energy component $G_3(k)$ of our arousal model as follows:

$$G_3(k) = 100 e_n(k)(1 - \overline{e}_n)\%$$
$$\text{where } e_n(k) = \frac{\widetilde{e}(k)}{\max_k(\widetilde{e}(k))} \text{ and } \overline{e}_n = \frac{1}{W}\sum_k e_n(k). \quad (6)$$

4) Arousal Model: Fig. 9 shows all three arousal components computed for an excerpt from a soccer match. When compared with the content description of characteristic segments of this excerpt (see labels), one can see that at the times of exciting events (goals, chances), local maxima exist in all three curves, as opposed to less exciting segments. One can also notice that these local maxima are not necessarily aligned. For instance, in the case of a score, the following scenario is possible: the spectators first cheer the action (sound energy peak), then there are cameras zooming to running players (motion activity peak) and, finally, there are cameras zooming to the teams' benches and to spectators (cut density peak). This fact motivates the definition of the function F as a weighted average of the three components, which is then convolved with a sufficiently long smoothing window in order to merge neighboring local maxima of the components. The result is finally re-scaled to the 0%–100% range. The process is shown in (7), as follows:

$$A(k) = \frac{\max_k(a(k))}{\max_k(\widetilde{a}(k))}\widetilde{a}(k)\%$$
$$\text{with } a(k) = \sum_i w_i G_i(k) \text{ and } \widetilde{a}(k) = K(l_2, \beta_2) * a(k). \quad (7)$$

Here, w_i are the coefficients weighting the component functions $G_i(k)$ with $\sum_i w_i = 1$. For the purpose of smoothing, we again apply the Kaiser window. However, as indicated by the values l_2 and β_2, this window may have a different length and shape parameter compared to the window used previously for the three components.

5) Evaluation: Having described in detail the methods how to model the three arousal components, and to integrate the components as in (7) to form a complete arousal model, we now proceed to validate these using real media data. The choice of test video sequences was based on two considerations. First, in order to obtain meaningful results, the sequences selected should be those in which the changes in arousal are most likely induced by the stimuli depicted by the low-level features adopted. Second, the sequences selected should be characterized by the content flow on which an average user is expected to react in a "standard manner" in terms of arousal. For instance, the arousal is expected to rise when the development of a soccer game goes from the stationary ball exchange in the middle of the field and finishes with the score via a surprisingly forward push toward the goal. In the same fashion, the arousal is supposed to decrease with the stabilization of a situation in an action movie, following a rapid action event.

Our test set includes excerpts from two different soccer matches (from two different broadcasters, separately) as well as excerpts from the movies "Saving Private Ryan" and "Jurassic Park 3." We observed the behavior of the arousal time curves in the global sense, and checked whether it complies with the content development along various sequence segments. At the same time, we checked the similarity of the arousal levels obtained for similar events in different sequences. For each test sequence, the same set of parameter values has been used: pixel block size for motion estimation was selected as 16, the coefficients w_i were selected as 1/3, and δ was set to 300. The length and shape parameter of the Kaiser window used for arousal components were 700 and 5, and those for the complete arousal model were 1500 and 5, respectively.

Figs. 9 and 10 show the arousal time curves obtained for the test sequences. In each curve the characteristic segments are labeled to reveal the actual content of the corresponding sequence such that the model performance can be judged. By examining these results we can conclude that the arousal levels of similar events in different sequences (e.g., goals in soccer games) are comparable, and that the obtained distributions of arousal levels along each sequence correspond to expectations.

D. Model for Valence

The *Compatibility* criterion described in Section V-A requires that the affect curve generated through combining the arousal and valence time curves should cover an area in the valence-arousal coordinate system that has a parabolic-like shape resembling the 2-D emotion space (Fig. 3). Clearly, this

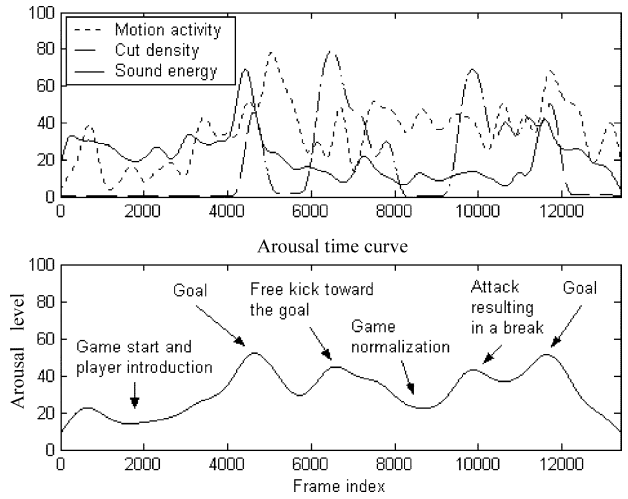

Fig. 9. Computation of the arousal time curve by superimposing the component time curves.

criterion confines that the values of arousal and the absolute values of valence are related to each other, which means that in general the range of arousal values determines the range of absolute valence values. We, therefore, start the development of the valence model by defining the function $r(k)$ that captures this value range dependence

$$r(k) = a(k) \cdot \text{sign}\{H(D_j(k), j = 1, \ldots, M)\} \qquad (8)$$

where, as usual, k is the frame index, and $a(k)$ as defined in (7) is the arousal function before smoothing. Similar to the discussion on the arousal model (1), each component $D_j(k)$ in (8) models the changes in valence as revealed by the feature j, while the function H serves to integrate the contributions of all the components in the final valence time curve. Clearly, the values $r(k)$ are determined solely by the values of the arousal, while the function H only determines the sign of $r(k)$.

The values of H are used in the next step to compute the variations of the valence in the value range specified by the arousal. In order for the valence values to remain in the proper range, the amplitude of these variations needs to be much smaller than the value of the arousal determining that range. With this in mind, we define the variance function $g(k)$ as follows:

$$g(k) = \frac{n}{100} \cdot \max_k A(k) \cdot \frac{H(D_j(k), j = 1, \ldots, M)}{\max_k |H(D_j(k), j = 1, \ldots, M)|}\%. \qquad (9)$$

The number n determines the magnitude of allowable variations of valence values in the range specified by the arousal. As shown in (9), this magnitude is not allowed to exceed n percent of the maximum arousal value.

We now model the valence time curve as

$$V(k) = \frac{\max_k |v(k)|}{\max_k |\widetilde{v}(k)|}\widetilde{v}(k)\% \quad \text{with}$$
$$v(k) = r(k) + g(k) \quad \text{and} \quad \widetilde{v}(k) = K(l_2, \beta_2) * v(k). \qquad (10)$$

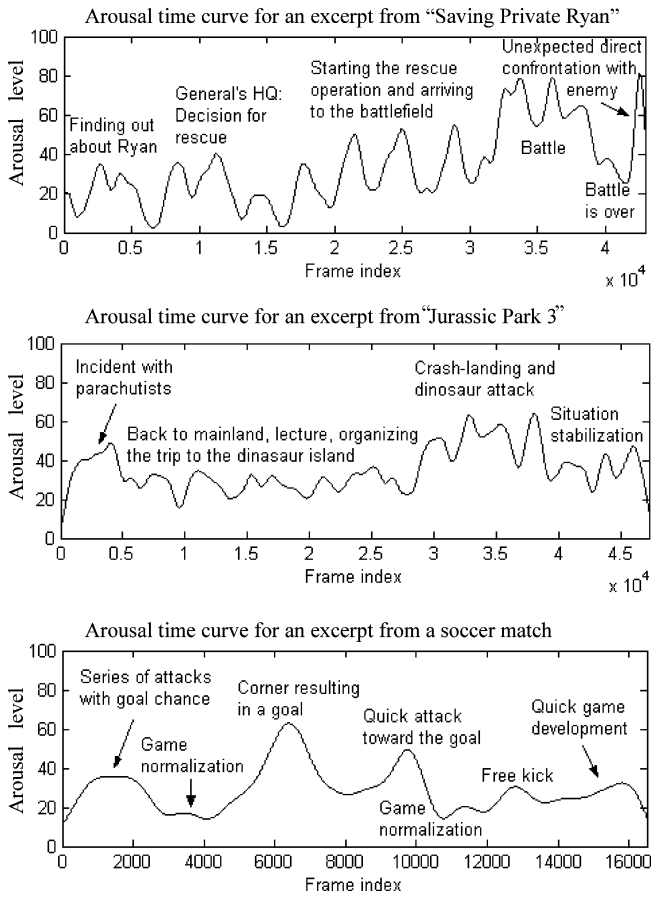

Fig. 10. Arousal time curves obtained for three test sequences. The labels describe the actual content of various sequence segments.

The smoothing window used here is the same as the one used for smoothing the final arousal time curve in (7). The main purpose of smoothing the curve $v(k)$ is to eliminate jumps appeared in the $r(k)$ curve due to the sign change in (8).

As is clear from discussions above, the role of function H is actually analogous to that of function F in (1). Therefore, the search for the proper form of function H can be done in the similar way as for function F. In the following, we first describe how to model a component function $D_j(k)$ using one of the valence-related features – the *pitch average* – such that it satisfies the criteria of comparability and smoothness. We then demonstrate the concept of modeling the valence time curve as explained above based on the example of the simple curve derived from the pitch-average component.

1) Pitch-Average Component: We compute the pitch signal using the off-the-shelf software and average the pitch values temporally over each video segment of length L. This results in the pitch-average time curve $P(k)$. As studied by Murray and Arnott [23], the average pitch can be useful in distinguishing between some positive and negative affective states, such as "happiness" (high-pitch average) and "sadness" (low-pitch average).

In order to associate the average pitch value with a corresponding valence value that may also be negative, we define the following function:

$$p(k) = P(k) - N. \qquad (11)$$

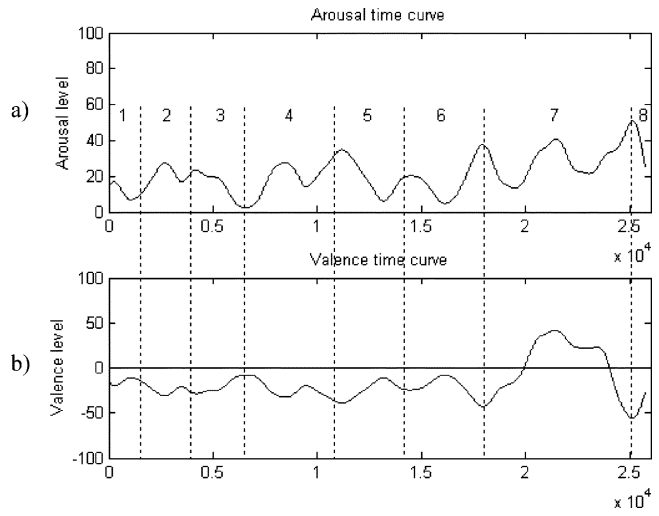

Fig. 11. (a) Arousal curve obtained for an excerpt from the movie "Saving Private Ryan." (b) Valence curve obtained for the same excerpt on the basis of the pitch-average component (12).

Here, N is what we call the "neutral feeling" frequency, and serves to map the low (high) values of the pitch average to the corresponding negative (positive) valence values.

In view of the smoothness criterion, the function (11) is not directly suitable to serve as a valence component time curve due to its step-wise nature. We therefore smooth the values (11) using the same Kaiser window as in the case of the arousal components. The result is the pitch-average component $D_1(k)$ of the valence time curve

$$D_1(k) = \frac{\max\limits_k |p(k)|}{\max\limits_k |\widetilde{p}(k)|}\widetilde{p}(k)\% \quad \text{with} \quad \widetilde{p}(k) = K(l_1, \beta_1) * p(k). \tag{12}$$

2) Evaluation: The measurement of the emotion type (valence) is much more ambiguous than the measurement of the emotion intensity (arousal). We therefore choose to evaluate the valence model (10) in its simplest form, where the function H is based on one component function only, that is, $H(D_j(k), j = 1, \ldots, M) = D_1(k)$, and in a controlled situation. The purpose of evaluation in this section is to prove the concept of

- modeling the valence components as shown by the example (12);
- modeling the valence time curve along the steps (8–10);
- generating the affect curve on the basis of the corresponding arousal and valence curves, as explained in Section IV-B.

Since we choose $D_1(k)$ as the pitch-average component, we select a test video sequence such that its emotional load can largely be determined on the basis of the pitch average only. For this purpose, we selected an excerpt from the movie "Saving Private Ryan" where the soundtrack consists of male voices that are only sporadically interrupted by noise or music.

Fig. 11 shows the arousal and valence time curve obtained for the selected test sequence. Besides the parameters already specified in Section V-C5, additional parameters here are the "neutral feeling" frequency N that is set to 150 Hz [26], the value of n that is set to 10%, and the pitch-average segment length L that is set to 909 frames. The rather odd value that we used for L resulted from our attempt to partition the test sequence into the segments of equal length, which are, at the same time, synchronized with the segments used to compute the pitch. In order to evaluate the correlation between the obtained arousal and valence values and the actual content of the sequence, we have labeled different parts of the sequence to describe their contents in as much detail as possible. These labels can be found in Table I.

Fig. 11(a) shows that the changes in arousal are not that strong. This was expected as the entire sequence is rather stationary and mainly contains conversations. A slight increase of the average arousal value in the segment 2, 4, and 7 as compared to the previous segment, is, however, quite correlated to the actual content development of the sequence along these segments.

The range of the valence values in Fig. 11(b) indicates that the valence time curve is basically a scaled (and mirrored, where negative) version of the arousal curve, on which the allowed variations modeled using the pitch-average component are superimposed. The first interesting spot in Fig. 11 is the switch of the valence curve from the negative to positive values around the frame 20 000. This switch reveals the change in the prevailing mood from mostly somber in the first part of the sequence to a "casual" every-day mood and even some happiness. This is then followed by, again, expected switch of the curve to the range of negative valence in the segment 8. The course of the obtained valence time curve largely corresponds to expectations. However, the simplicity of the function H has also lead to slight imperfections in the obtained curve. Namely, segments 5 and 6 also contain parts that are characterized by the similar "casual" every-day mood as in the segment 7. These parts are not properly revealed by the valence time curve in Fig. 11(b).

We now combine the arousal and valence curve from Fig. 11 in the affect curve that provides the complete representation of the affective content of the video clip under study. The parabolic shape of this curve shown in Fig. 12 clearly indicates the compatibility of the obtained curve with the 2-D emotion space. As we can read from the curve, the prevailing mood of the test sequence is rather somber (low-to-medium arousal and negative valence) with the exception of one segment that is characterized by a mid-level arousal and a positive valence.

VI. Discussion

In this paper, we first described the problem of extracting the affective content of an arbitrary video and revealed the basic scope of opportunities that would become realistic if a solution to this problem were found. The opportunities elaborated in Section II are within the context of video indexing and personalized video delivery. Then, we outlined the technique developed to extract and represent the affective content from a video, which has been motivated by studies in psychophysiology. After adopting a "dimensional approach to affect," that is, the possibility of representing the affective content using points in the 2-D emotion space, we have defined the links between dimensions of the emotion space and low-level features that can be extracted from video data using standard video and audio processing tools. As a result, we managed to obtain time curves

TABLE I
LABELS DESCRIBING THE CONTENT OF THE TEST SEQUENCE IN FIG. 12

Segment	Content description
1	US Army HQ, typists make letters for soldiers' families, male voices reading the letters
2	Colonel's office, finding out about private Ryan
3	Bad news brought to Ryan's home
4	General's office, decision is being made to search for Ryan
5	Omaha Beach, US Army HQ, an officer gets the order to search for Ryan
6	Omaha Beach, US Army HQ, preparation for the search action
7	Beginning of the action, walking through the fields
8	It starts to rain and gets dark, the suspense grows, the actual beginning of the action

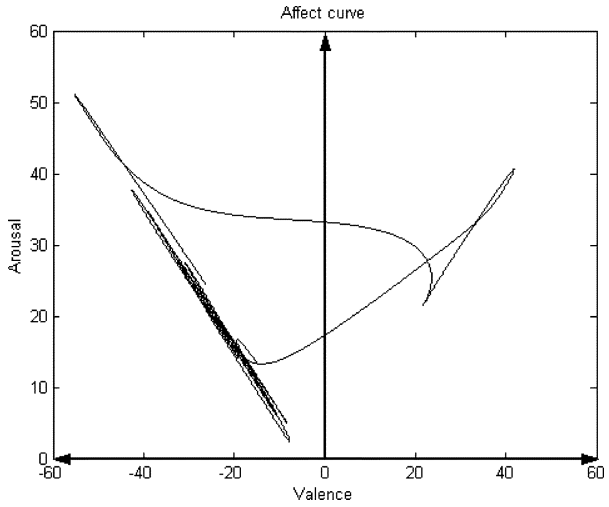

Fig. 12. Affect curve obtained by combining the curves from Fig. 11(a)–(b).

that represent, respectively, the expected transitions from one arousal and valence level to another along a video, as perceived by a viewer. For some video genres, such as movies, it is also advisable to combine the obtained time curves into the so-called affect curve that can serve to determine the prevailing mood per segment of a video or of the video in its entirety.

We tested our models on a number of video sequences belonging to different genres. The results are largely in line with the expectation, concerning both separate arousal and valence curves and the final affect curve. Although some of the tests were performed in a controlled situation, we can claim that the proposed arousal and valence models represent a solid basis for obtaining a reliable affective video content representation. We see the possibility for further improvement of the obtained representation in searching for more concrete relations between the affect dimensions (arousal and valence) and low-level features. As can be seen from the discussion in Section V-B, the relations known so far are rather vague and therefore difficult to map onto reliable models for arousal or valence components.

REFERENCES

[1] B. Adams, C. Dorai, and S. Venkatesh, "Novel approach to determining tempo and dramatic story sections in motion pictures," in *Proc. IEEE ICIP 2000*, vol. II, Vancouver, BC, Canada, 2000, pp. 283–286.
[2] R. Arnheim, *Film as Art*. London, U.K.: Faber & Faber, 1958/1983.
[3] N. Babaguchi, "Toward abstracting sports video by highlights," in *Proc. IEEE ICME 2000*, vol. 3, pp. 1519–1522.
[4] D. Bordwell and K. Thompson, *Film Art: An Introduction and Film Viewers Guide*, 7th ed. New York: McGraw-Hill, 2003.
[5] M. Bradley, "Emotional memory: A dimensional analysis," in *Emotions: Essays on Emotion Theory*. Hillsdale, NJ: Lawrence Erlbaum, 1994.
[6] M. M. Bradley and P. J. Lang, *International Affective Digitized Sounds (IADS): Technical Manual and Affective Ratings*. Gainesville, FL: Center for Res. Psychophysiol., Univ. Florida, 1991.
[7] A. Del Bimbo, *Visual Information Retrieval*. New York: Morgan Kaufmann, 1999.
[8] B. H. Detenber, R. F. Simons, and G. G. Bennett, "Roll 'em!: The effects of picture motion on emotional responses," *J. Broadcasting and Electron. Media*, vol. 21, pp. 112–126, 1997.
[9] R. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Proc. Cognitive Technology Conf.*, San Francisco, CA, 1999.
[10] L. Fitzgibbons and R. F. Simmons, "Affective response to color-slide stimuli in subjects with physical anhedonia: A three-systems analysis," *Psychophysiology*, vol. 29, no. 6, pp. 613–620, 1992.
[11] L. D. Giannetti, *Understanding Movies*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1976.
[12] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *J. Psychophysiol.*, vol. 3, pp. 51–64, 1989.
[13] A. Hanjalic, G. C. Langelaar, P. M. B. Van Roosmalen, R. L. Lagendijk, and J. Biemond, *Image and Video Databases: Restoration, Watermarking and Retrieval*. Amsterdam, The Netherlands: Elsevier Science, 2000.
[14] R. Hopkins and J. E. Fletcher, "Electrodermal measurement: Particularly effective for forecasting message influence on sales appeal," in *Measuring Psychological Responses to Media*, A. Lang, Ed. Hillsdale, NJ: Lawrence Erlbaum, pp. 113–132.
[15] A. Jaimes, T. Echigo, M. Teraguchin, and F. Satoh, "Learning personalized video highlights from detailed MPEG-7 metadata," in *Proc. IEEE ICIP*, vol. 1, 2002, pp. 133–136.
[16] A. Lang, P. Dhillon, and Q. Dong, "Arousal, emotion, and memory for television messages," *J. Broadcasting and Electron. Media*, vol. 38, pp. 1–15, 1995.
[17] A. Lang, J. Newhagen, and B. Reeves, "Negative video as structure: Emotion, attention, capacity, and memory," *J. Broadcast. Electron. Media*, vol. 40, pp. 460–477, 1996.
[18] P. J. Lang and M. Bradley, "Measuring emotion: The self-assesment Manikin and the semantic differential," *J. Behavior Therapy & Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[19] P. J. Lang, *The Network Model of Emotion: Motivational Connections*, R. S. Wyer and T. K. Srull, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1995, vol. 6, Advances in Social Cognition.
[20] P. J. Lang, A. Öhman, and D. Vaitl, "The International Affective Picture System [Photographic Slides] Tech. Rep.," Center for Res. in Psychophysiol., Univ. Florida, Gainsville, 1988.
[21] *Principles of Visual Information Retrieval*, M. Lew, Ed., Springer-Verlag, Berlin, Germany, 2001.
[22] B. Li and M. I. Sezan, "Event detection and summarization in American football broadcast video," in *SPIE/IS&T Storage and Retrieval for Media Databases 2002*, 2002, pp. 202–213.
[23] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, Feb. 1993.
[24] C. Osgood, G. Suci, and P. Tannenbaum, *The Measurement of Meaning*. Urbana, IL: Univ. Illinois Press, 1957.
[25] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
[26] R. Picard and G. Cosier, "Affective intelligence – The missing link?," *BT Technol. J.*, vol. 14, no. 4, pp. 150–161, October 1997.

[27] D. Romer, "The Kodak Picture exchange," in *Seminar at MIT Media Lab*, Cambridge, MA, Apr. 1995.
[28] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, pp. 273–294, 1977.
[29] R. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, "Emotion-processing in three systems: The medium and the message," *Psychophysiology*, vol. 36, pp. 619–627, 1999.

Alan Hanjalic (M'00) received the Dipl.-Ing. degree from the Friedrich-Alexander University, Erlangen, Germany, in 1995 and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1999, both in electrical engineering.

Currently, he is a tenured Assistant Professor at the Department of Mediamatics, Delft University of Technology. He was a Visiting Scientist at Hewlett-Packard Labs, Palo Alto, CA, in 1998, a Research Fellow at British Telecom Labs, Ipswich, U.K., in 2001, and a Visiting Scientist at Philips Research Labs, Briarcliff Manor NY, in 2003. His research interests are in the broad areas of multimedia signal processing, media informatics and multimedia information retrieval, with focus on multimedia content analysis for interactive content browsing and retrieval and on personalized and on-demand multimedia content delivery. In his areas of expertise, he authored and coauthored more than 40 publications, among which the books titled *Image and Video Databases: Restoration, Watermarking and Retrieval* (Amsterdam, The Netherlands: Elsevier, 2000) and *Content-Based Analysis of Digital Video* (Norwell, MA: Kluwer, 2004)

Dr. Hanjalic was a Guest Editor of the Special Issue on Content-based Image and Video Retrieval of the *International Journal of Image and Graphics* (July 2001) and was the initiator and main organizer of the Symposium on Multimedia Retrieval, Eindhoven, The Netherlands, in January 2002. He regularly serves either as a Program Committee member, Session Chair, or a panel member in the IEEE and SPIE/IS&T conferences. He is an advisor/reviewer of the Belgian Science Foundation (IWT) in the area of information technology and systems. Since 2002, he has been the Secretary of the IEEE Benelux Section. He is also a member of the Organizing Committee of the IEEE International Conference on Multimedia and EXPO (ICME) 2005, to be held in Amsterdam.

Li-Qun Xu (M'03) received the Ph.D. degree in information engineering from Southeast University, Nanjing, China, in 1988.

He joined BT Research and Venturing, Ipswich, U.K., in 1996 as a Senior Researcher, where he is currently a Principal Researcher and Project Manager in the Broadband Applications Research Centre. His recent research interests are in the broad areas of Visual Information Processing, including multimedia content analysis and indexing, robust object segmentation and tracking for intelligent visual surveillance, people behaviour and event analysis, 2-D motion analysis and segmentation, 3-D vision techniques and image-based rendering for collaborative working environment, among others. He has published prolifically on these and allied topics and holds a number of patents and pending applications. Prior to his career with BT, he has was an academic in a number of British Universities, both as a member of the research staff and lately within the faculty between 1990 and 1996.

Dr. Xu is a member of British Computer Society and a member of IEEE Signal Processing and Computer Societies.