Adaptive Extraction of Highlights From a Sport Video Based on Excitement Modeling

Alan Hanjalic, *Member, IEEE*

***Abstract*—This paper addresses the challenge of automatically extracting the highlights from sports TV broadcasts. In particular, we are interested in finding a generic method of highlights extraction, which does not require the development of models for the events that are thought to be interpreted by the users as highlights. Instead, we search for highlights in those video segments that are expected to excite the users most. It is namely realistic to assume that a highlighting event induces a steady increase in a user's excitement, as compared to other, less interesting events. We mimic the expected variations in a user's excitement by observing the temporal behavior of selected audiovisual low-level features and the editing scheme of a video. Relations between this noncontent information and the evoked excitement are drawn partly from psychophysiological research and partly from analyzing the live-video directing practice. The expected variations in a user's excitement are represented by the excitement time curve, which is, subsequently, filtered in an adaptive way to extract the highlights in the prespecified total length and in view of the preferences regarding the highlights strength: extraction can namely be performed with variable sensitivity to capture few "strong" highlights or more "less strong" ones. We evaluate and discuss the performance of our method on the case study of soccer TV broadcasts.**

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

I. Introduction

With the advent of digital video[1] revolution the television broadcasting industry is slowly but surely transferring to the end-to-end digital television production, transmission and delivery chain. Supported by the availability of broadband communication channels, this transfer will lead to an enormous increase in the amount of video data reaching our homes. At the same time the quickly growing capacity-versus-price ratio of digital storage devices is likely to make such devices highly popular with consumers. A combination of the abovementioned phenomena will result in an explosion in the "consumer choice", that is, in the number of video hours that are instantaneously accessible to the consumer. This may have crucial consequences for the ways the broadcasted material is "consumed". As reported in the study of Durlacher Research Ltd. [1], the understanding of the broadcasting mechanism may change. This mechanism will only be something that provides data to the—soon inevitable—*home mass storage system* (HMSS), and as far as the consumer is concerned, the concept of the "broadcasting channel" will lose its meaning. Further, due to large amounts of incoming data, video recording will be performed routinely and automatically, and programs will be accessed on demand from the local storage. Viewing of live TV is therefore likely to drastically diminish with the time [1].

Various scientific and technological challenges appear as a consequence of the development described above, aiming at maximum transparency of the recorded video volume toward the consumer—independent of the volume size. One of such challenges is to develop techniques for automatically abstracting video. The purpose of a video abstraction algorithm in the context of an HMSS can be twofold. First, such an algorithm can be designed to *summarize* a TV program in order to facilitate the consumer's choice of what to watch later on. This may be highly valuable, for instance, in the process of digesting a large volume of news television broadcasts and presenting to the consumer in a compact but comprehensive way the coverage of all news topics found in the volume. Alternatively, a video abstraction algorithm can be designed to *prune* the recorded video material by keeping the most interesting segments—or *highlights*–only, and by leaving out the remaining, less interesting parts. In this paper we concentrate on the design of a video abstraction algorithm for pruning purposes, and address the specific problem of pruning sports television broadcasts. The sports programs are particularly interesting objects for pruning as they lack story line, and as the events being worth watching (e.g., goals in soccer, home runs in baseball, touchdowns in football) are sparse and spread over a long period of time.

The challenge of automatically pruning sports television broadcasts has been pursued widely in the scientific community in recent years [15]. An analysis of the previous work on this subject reveals that most of the approaches proposed so far are event-based: they aim at detecting predefined events that are considered most interesting in a particular sport genre. Event detection is approached either by developing feature-based event models [8], [12], [13], [22]–[24], [26], by searching for keywords in speech (e.g. commentator) [4] and closed captions [16], by using MPEG-7 metadata [11] or by involving several of the above-mentioned clues into inter-modal collaboration [3], [9], [21]. We see the main disadvantage of this approach in the need for numerous and reliable event models which should take into account not only all highlight-related events but also various realizations of these events and their coverage that may change from one broadcaster to another. Clearly, this makes the

Manuscript received October 6, 2004; revised September 21, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chitra Dorai.

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

Digital Object Identifier 10.1109/TMM.2005.858397

[1]We refer to "video" as a multimodal data stream consisting of an image sequence and the accompanying audio.

event-based highlight extraction technically and semantically a complex task in many broadcasts. Although the problem of different event realizations and coverage can partly be solved via keyword spotting, the composition of the highlighting video abstract in that case is limited to those events only, for which obvious keywords are likely to be found. While this may be the case for the "goal" and "penalty" events in soccer or "home run" in baseball, other interesting events, such as a nice action on the net during a tennis game or a quick attack toward the goal finishing by a nice move of a goalkeeper in soccer, will probably remain undetected.

An alternative to the approaches discussed above is to search for a single effect that is assumed to accompany an arbitrary highlighting event. For instance, Pan *et al.* [17] based the detection of highlights on the detection on slow-motion segments. They observe namely that the interesting events are often replayed in slow motion immediately after they occur. Although being more generic than the methods discussed above, this highlights extraction mechanism may result in a large number of falsely extracted video segments: interesting events are often replayed in slow-motion during several later game breaks. Further, several slow-motion segments may be played after each other, showing the highlighting event from different camera views. Similar problems emerge in the attempts to detect specific audio events, such as "applause", "cheering crowd" or "cheering commentator", that are often seen as indicators of the potential presence of a highlight. Although rather successful attempts to extract highlights based on an analysis of audio events alone were reported (e.g., [25]), Rui *et al.* [20] observe that such an approach, in general, is likely to lead to a large number of false alarms. The spectators and the commentator may, namely, become "loud" for the reasons that have nothing to do with the sport event considered (e.g., cheering a hot-air balloon that flies over the stadium). It is, therefore, not surprising that many proposed methods combine the "cheering" detection with other, in most cases event-specific clues [4], [5], [18], [20].

A further step toward the development of generic tools for sport-program pruning is made by the approaches for detecting the high-level sport program structure. For instance, Li and Sezan [14] develop both a deterministic and probabilistic framework for detecting the "play" segments of a sport video. These segments are assumed to contain all the interesting material of a program, as opposed to less interesting "nonplay" segments. They demonstrate the applicability of their framework to football, baseball and sumo wrestling. An alternative approach to detecting "play" and "break" events was proposed by Ekin and Tekalp [7], and was demonstrated on basketball, football, golf and soccer. However, although being very helpful in filtering out irrelevant video segments, these structure-oriented approaches are not suitable for highlights extraction: not all the material contained in a "play" segment can be considered a highlight. Further, the detection of structural elements as described above may not be possible in some sport disciplines, such as swimming or car racing.

As a spin-off from our previous work on affective video content representation and modeling [10], we propose in this paper a method for highlights extraction that is based on modeling the expected variations in the excitement level of the user while watching a sports video. Since it is realistic to assume that a highlighting event (e.g. goal, touchdown, home run, the finals of a swimming competition, or the last 50 meters in a running contest) induces a steady increase in a user's excitement, we search for highlights in those video segments that are expected to excite the user most. At this stage it is worthwhile emphasizing that the *expected* level of excitement should not be mixed up with the *actual* level of excitement 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 program director, or reflects the more-or-less unanimous response of a general audience to a given audio-visual stimulus. Opposed to this, the actual level of excitement is highly subjective and context-dependent. For instance, the same soccer television broadcast may evoke different levels of excitement at the winning team's and the losing team's fans, and elicit no reaction at all from an audience that is not interested in soccer. As we argue in [10], the expected variations in a user's excitement induced by video can be modeled as a function of various content-unrelated video features. In this way, the basis is provided for developing generic methods for highlights extraction independent of the type of events to be expected in a particular sports program genre but also independent of the event realization and coverage.

In Section II, we first recap the main points of our approach to excitement modeling from [10]. This approach generates an excitement time curve through a superposition of a number of excitement components, each of which mimics the changes in a user's excitement as a reaction to one particular audio-visual stimulus. Then, in Section III, we explain how a filtered version of the excitement time curve can be obtained, revealing only those video segments that contain an event of a particular "strength". Here, the strength is determined by the total excitement level evoked by an event and by the "richness" of event experience. Subsequently, the highlights are extracted in a predefined length L by cutting-off the peaks of the filtered excitement time curve. By tuning the "strength" parameter, the filtering can be performed adaptively. In this way, the highlights can be extracted with variable sensitivity to capture few "strong" highlights or more "less strong" ones by a constant abstract length L. In Section IV, we evaluate and discuss the performance of our method on the case study of soccer TV broadcasts. Conclusions and recommendations for future research are given in Section V.

II. Excitement Modeling

We introduced in [10] the *arousal time curve* $A(k)$ as a representation of the expected variations in a user's excitement while watching a video. Formally, the curve $A(k)$ was defined as a function of N basic components $G_i(k)$

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

To simplify the terminology, we will refer to $A(k)$ in this paper as *excitement time curve*. The function $G_i(k)$ models the variations in the excitement level over the video frames k as induced by the stimulus represented by the feature i. We assume that all of the N features considered in the model (1) were selected to reliably represent the stimuli that will likely influence the affective state of the user while watching a video. For instance,

psychophysiological experiments have shown that the level of excitement of an average user is expected to rise as a consequence of an increase in sound energy and (camera/object) motion intensity in the video being watched [6], [19]. In this sense, each function $G_i(k)$ can be seen as an elementary excitement time curve or the primitive of the overall excitement time curve $A(k)$, while $A(k)$ is obtained by integrating the contributions of all components $G_i(k)$ using a suitable function F.

As excitement is a psychological category, we introduced in [10] two basic criteria that are to be satisfied by the models for both the $A(k)$ and $G_i(k)$ in order for these models to be psychologically justifiable. The first criterion (*Comparability*) ensures that the levels of excitement computed for similar events in different videos are comparable. This criterion obviously imposes normalization and scaling requirements when computing the abovementioned time curves. The second criterion (*Smoothness*) accounts for the degree of memory retention of preceding frames and shots. It ensures that the perception of the content, and consequently, of 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).

III. Adaptive Filtering of the Excitement Time Curve

Given the excitement time curve $A(k)$ and the maximum desired abstract length L in frames, the simplest approach to highlights extraction is to look at the values of the curve and to extract those video segments that are likely to excite the user most. To do this, we can draw a horizontal line cutting off the peaks of the curve in such a way that the number of frames in video segments where the peaks are found is not larger than L [10]. As the extraction process is driven by the local excitement level only, any event in a sport program may be included into the abstract, provided that the curve $A(k)$ passes through a sufficiently high value range during that event. In this way, highlights are extracted in a generic fashion without the need for event modeling or artificially limiting the scope of the abstract content.

In view of the fact that each value of the curve $A(k)$ results from multiple combined stimuli represented by the component time curves $G_i(k)$, we could also extract the highlights in a more sophisticated fashion, namely by taking into account the additional criterion of highlight "strength". We define the strength of a highlight as the number M of component time curves $G_i(k)$ that rise to a high value range during this highlight. The number M is therefore equivalent to the number of stimuli that have major influence on the affective state of the user during a particular sport event, and can be said to determine the "richness" of the experience of that event: the richer the experience the stronger (better, more interesting) is the highlight. To ensure the selectiveness of the highlights extraction process, given the minimum allowed highlight strength M only those video segments having the strength of at least M should be allowed to enter the process in the first place. This can be done by "filtering" the original excitement time curve $A(k)$: the curve values in video segments that are likely to be the highlights of the required strength are left high while all other curve values are pulled down in order not to be captured by the cutoff line.

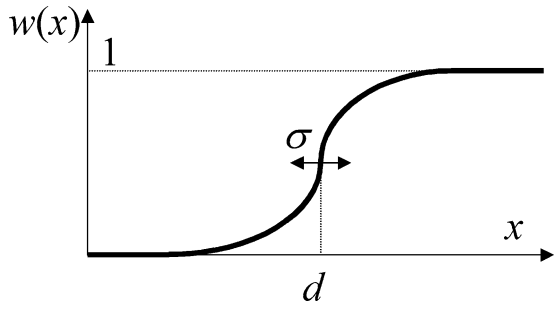

Fig. 1. Weighting function $w(k)$.

Let us consider the time stamp k and the values of the components $G_i(k)$ computed at that time stamp. We rank the values $G_i(k)$ in the descending order, denote the ranked values by $\overline{G}_j(k)$ and consider the elements of the subset $S_M = \{\overline{G}_j(k) | j = 1, \ldots, M\}$. As the lower bound of the value range of all components in the subset S_M is determined by the value of the minimum of that subset, $\overline{G}_M(k)$, we weight the values of the curves $G_i(k)$ in correspondence to this minimum, that is

$$G'_i(k) = G_i(k) w(k), \quad i = 1, \ldots, N \tag{2a}$$

where

$$w(k) = \frac{1}{2}\left(1 + erf\left(\frac{\overline{G}_M(k) - d}{\sigma}\right)\right) \tag{2b}$$

and

$$erf(x) = \frac{2}{\pi}\int_0^x e^{-t^2} dt. \tag{2c}$$

The behavior of the weighting function $w(k)$ is illustrated by the curve in Fig. 1. The parameters d and σ are the delay from the origin and the spread factor determining the steepness of the middle curve segment, respectively. The form of the weighting function (2b) secures that the weighted result is as discriminative as possible while, at the same time, no reasonable highlight candidate is discarded, as it would be the case if a hard threshold was applied to evaluate the argument in (2b) instead. If the minimum of the subset S_M is not high enough, the weighting factor $w(k)$ is close to zero. Consequently, the components $G_i(k)$ are pulled down, and therewith also the resulting excitement value $A(k)$. This is not the case only at those time stamps where the value $\overline{G}_M(k)$ is sufficiently high, meaning that the values of all components from the subset S_M are sufficiently high. There, the value of $w(k)$ is close to one and the effect of filtering is negligible. Clearly, the filtering process is adaptive, as it is controlled by the parameter M: the higher the value of M, the more strict is the filtering of the excitement time curve.

By considering the processed components $G'_i(k)$ in the model (1), instead of the original components $G_i(k)$, we obtain a filtered version of the excitement time curve that we will refer to as *highlights time curve* $H_M(k)$

$$H_M(k) = F\left(G'_i(k), i = 1, \ldots, N\right). \tag{3}$$

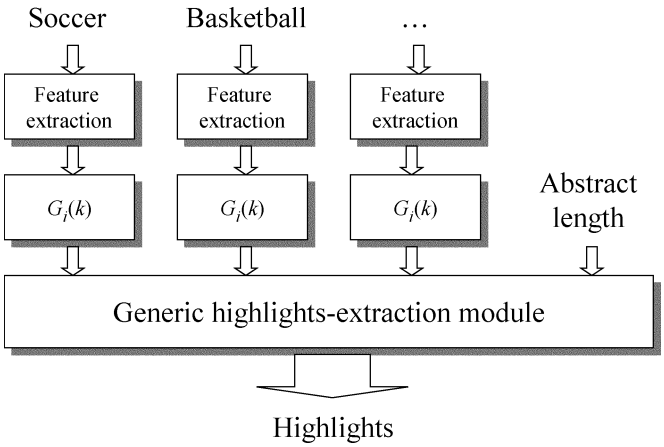

Fig. 2. Scheme illustrating the possibilities for practical implementation of the proposed highlights extraction method.

The time curve $H_M(k)$ serves, instead of the curve $A(k)$, as the basis for extracting highlights using the methodology explained in the first paragraph of Section III. By applying the cutoff line to the highlights time curve $H_M(k)$ only those video segments will be considered highlights in which the excitement values remain high after the filtering process (2). Consequently, the parameter M controlling the filtering process can also be said to determine the composition of the highlighting video abstract.

Fig. 2 illustrates the possibilities for practical implementation of the proposed adaptive highlights-extraction method. While the filtering process (2), the formation of the highlights time curve and the actual highlights extraction using a cutoff line can be considered universal for all sports program genres, this is not necessarily true for the feature set used to compute the component time curves $G_i(k)$. This is simply due to the fact that some features revealing the excitement stimuli in one genre may not be present in another genre or may not be relevant for excitement measurement in that genre.

IV. A Case Study: Soccer

A. Creation of a Highlights Time Curve

In order to evaluate our proposed method for highlights extraction, we choose the "soccer" program genre and the realization of the excitement model (1) that we proposed in our previous work [10]. This realization includes an expression for the function F and a methodology for computing the component time curves $G_i(k)$ for the following excitement-related features.

(a) the overall motion activity measured at frame transitions;
(b) the rhythm, obtained by investigating the changes in shot lengths along the video;
(c) the energy contained in the audio track of a video.

To the best of our knowledge, this excitement model realization is the only one proposed in literature so far. Furthermore, the features mentioned above can intuitively be related to many events in a soccer television broadcast that are generally considered "exciting". As explained in [10], each of these features influences the level of excitement in the positive way: an increase in a feature value leads to an increase in the excitement level of the user. We can illustrate the applicability of these features to our "soccer" case study on the example of a typical "exciting" event in a soccer match—a goal. First, we can expect a sound energy peak produced by the audience or a commentator cheering the action. The sound energy is much stronger there compared to the (stationary) segments of the match surrounding the goal. Further, the "goal" event is likely to be characterized by the motion activity peak that results from extreme motion activity in close-up shots of running players celebrating the goal. Finally, there may be a peak in the rhythm that results from a strongly increased shot change rate produced by the director to illustrate the atmosphere in the stadium after the goal (e.g. short close-ups to the bench, spectators, players, etc.).

The features mentioned above have already been employed in various contexts directly or indirectly related to the objectives of this paper. For instance, motion magnitude and shot lengths were used in [2] to model the "pace" function of a movie. Further, the sound energy was used by many authors for the purpose of video abstraction, as already reported in the literature overview in Section I. For modeling the component time curves $G_i(k)$, however, we employ these features in a different way [10] in order to fulfill the requirements of *Comparability* and *Smoothness*. The obtained components $G_i(k)$ are subsequently processed by the (2), which results in the modified component time curves $G_i'(k)$. According to the definition of the integrating function F in [10], we base the highlights time curve $H_M(k)$ on a weighted average of the modified components $G_i'(k)$

$$H_M(k) = \frac{\max\limits_k (a(k))}{\max\limits_k (\widetilde{a}(k))} \widetilde{a}(k)\% \tag{4a}$$

with

$$a(k) = \sum_i \eta_i G_i'(k) \quad \text{and} \quad \widetilde{a}(k) = K(l,\beta) * a(k). \tag{4b}$$

Here, η_i are the coefficients weighting the component functions $G_i'(k)$, with $\sum_i \eta_i = 1$. The convolution introduced in (4b) appears particularly useful to merge the neighboring local maxima into one "highlight peak" and so to compensate for the possible (and realistic) slight asynchrony of the behavior of the component time curves at places of highlights. The convolution is performed with the Kaiser window of the length and shape parameter l and β, respectively. Fig. 3 shows an example of the highlight time curve $H_M(k)$ without the smoothing step [Fig. 3(a)], and after the smoothing has been applied [Fig. 3(b)]. In case of no smoothing, the extracted abstract for the event (goal) in the time interval between the frames 6000 and 6800 will consist of two short segments, while after smoothing the entire goal event is extracted as one segment.

B. Highlights Extraction

We first demonstrate the soccer highlights extraction procedure described above on the example of one excerpt from a typical soccer television broadcast. The length and shape parameter of the Kaiser window used to smooth the highlights time curve were set to 1500 and 5, respectively. The values d and σ in (2b) were selected as 40 and 5, respectively. Fig. 4(a) shows the time curves of the three defined excitement components obtained for this video sequence. These components served as the basis for

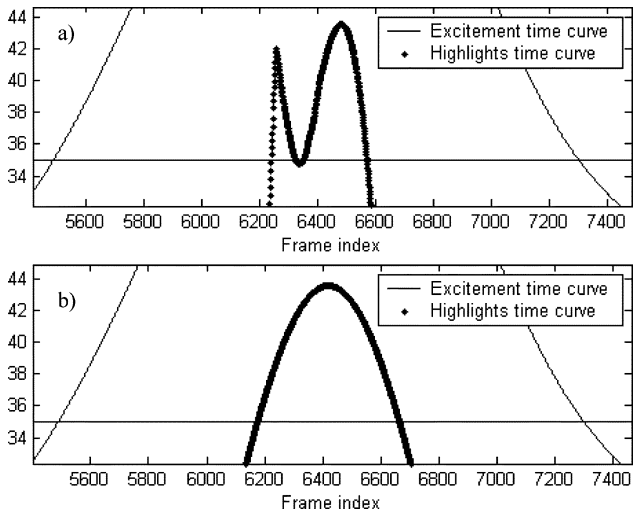

Fig. 3. Zoom-in on the excitement time curve, horizontal cutoff line for highlights extraction and the highlights time curve for the case of: (a) no smoothing and (b) smoothing of the weighted average of the component time curves.

computing the highlights time curve $H_M(k)$ for three possible values of $M = 1$, 2, or 3.

Let us first consider the case of $M = 3$, where we are maximally selective when creating the highlighting video abstract: we are interested in extracting the strongest (richest) highlights only. The highlights time curve for $M = 3$ is shown in Fig. 4(b) together with the original (nonfiltered) excitement time curve $A(k)$ that is obtained using the realization of the model (1) from [10]. If we look at the content labels characterizing different video segments in Fig. 4(a), we can see that the highlights time curve in Fig. 4(b) provides highly distinguishable peaks at video segments corresponding to goals. Obviously, the goals appear to be the only events in the analyzed segment of the soccer TV broadcast complying with the requirements that we posed on the strength of the highlights. The horizontal line in Fig. 4(b) provides an abstract of 50 s, showing the two goals contained in the analyzed excerpt, each of them preceded by the action leading to the goal and succeeded by a number of shots showing the situation in the stadium, as well as by replays of the action taken from different camera views. The effect of filtering becomes clear when we move the cutoff line vertically: depending on the line position, only the length of the extracted video segments will change, but not the composition of the resulting highlighting video abstract: it will always consist only of the goals and the actions related to them, while all other events of the game will be left out.

Fig. 4(c) and (d) show the highlights time curves obtained using the same procedure as above but with weaker requirements posed on the strength of the highlights to be extracted, namely, with $M = 2$ and $M = 1$, respectively. Clearly, each reduction of the value of M resulted in an extension of the scope of extractable video content. In Fig. 4(c), besides the goals also a goal chance (free kick) and a quick action toward the goal stopped by a game break are now considered in the highlights extraction procedure. In Fig. 4(d), no further events are added to the highlights extraction base, except that more material is considered related to the "free kick" event between frames 6000 and

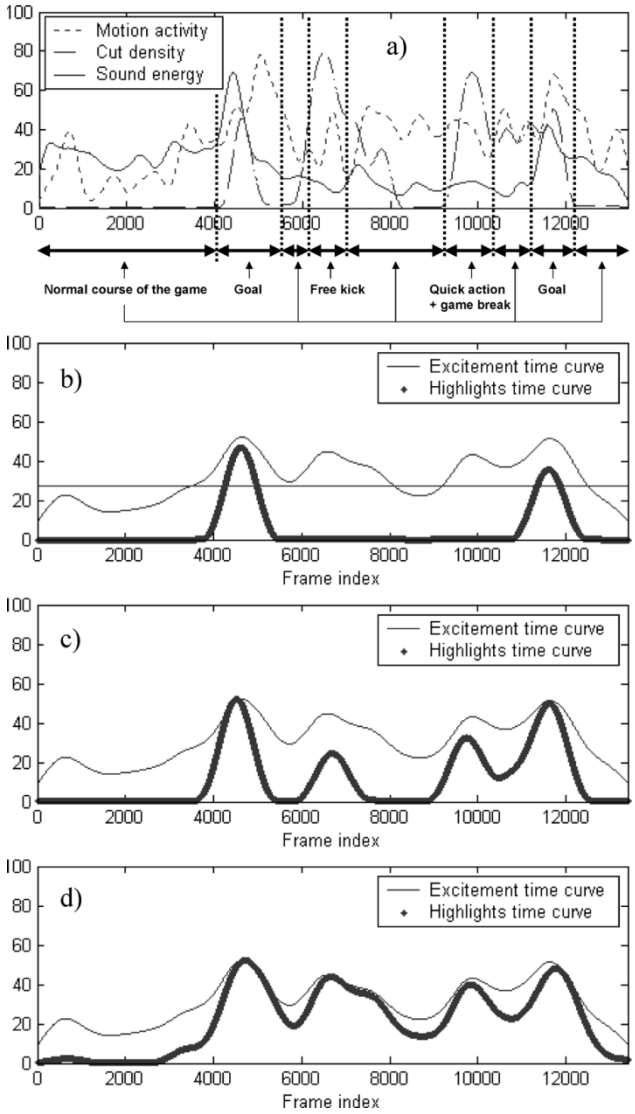

Fig. 4. (a) Component time curves obtained for an excerpt from a soccer match. Excitement and highlights time curve obtained for the case of (b) $M = 3$, (c) $M = 2$, and (d) $M = 1$. Horizontal cutoff line in (b) serves to extract highlights.

8000. This was also expected as in this video sequence no additional events could be found which could in one way or another qualify as highlights. The rest of the sequence consisted namely of more or less dynamic ball exchanges in the middle of the field which we referred to as the "normal course of the game". However, although the number of the extractable events is similar for $M = 2$ and $M = 1$, the ratio of the material extracted from different events will vary in both cases. As the peaks of the middle two events in Fig. 4(c) are lower than the peaks of the goals, much longer segments will be extracted for the goals relatively to other two events. This is not the case in Fig. 4(d), where the peaks of all four events have more equal height.

C. Discussion on Performance

We now extend our experimental evaluation to three other test sequences and discuss the robustness of the proposed highlights extraction method. Figs. 5–7 show the highlights time curves obtained for the new test video sequences, for all three values

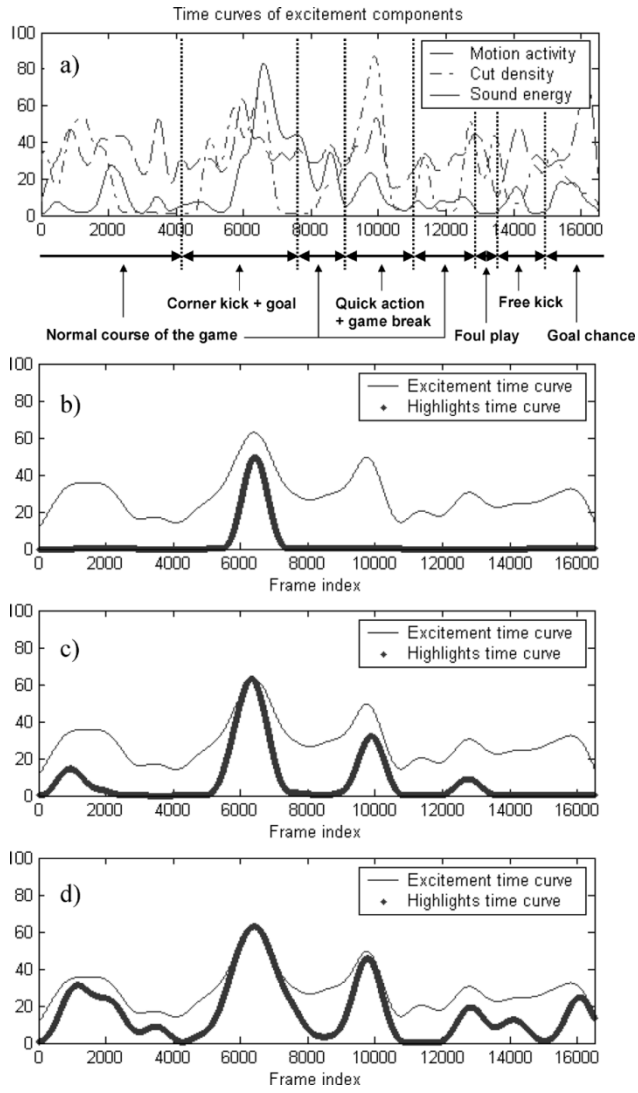

Fig. 5. (a) Component time curves obtained for an excerpt from a soccer match. Excitement and highlights time curve obtained for the case of (b) $M = 3$, (c) $M = 2$, and (d) $M = 1$.

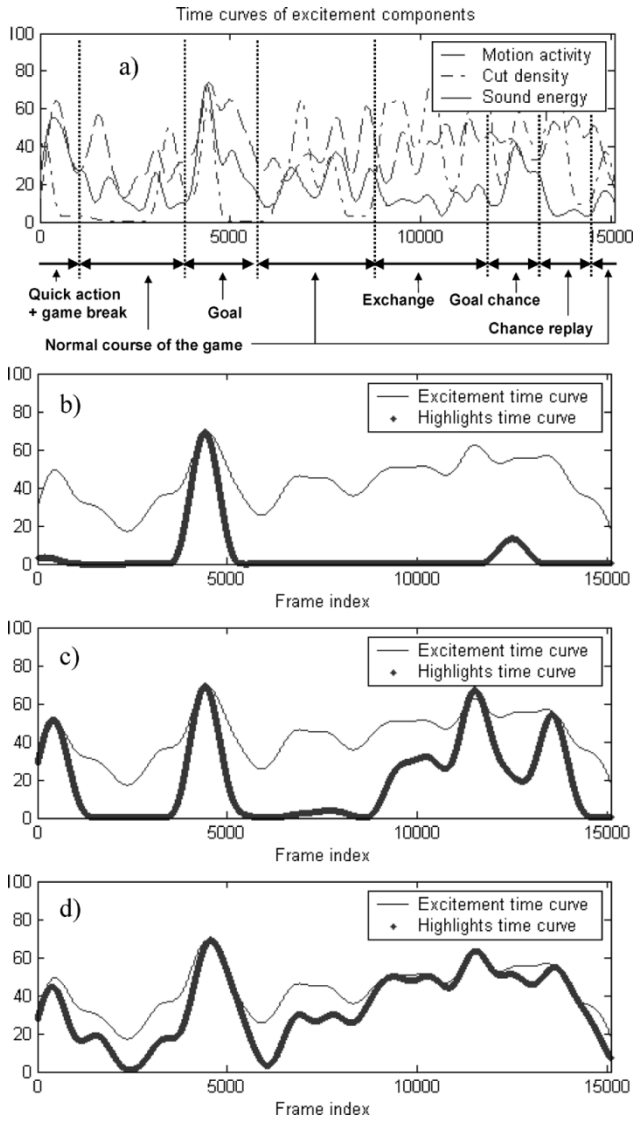

Fig. 6. (a) Component time curves obtained for an excerpt from a soccer match. Excitement and highlights time curve obtained for the case of (b) $M = 3$, (c) $M = 2$, and (d) $M = 1$.

of M and with the same parameter set as in Fig. 4. The test sequences belong to two different soccer television broadcasts produced by two different broadcasters, a Spanish (Figs. 4 and 7) and a British (Figs. 5 and 6) one.

1) The Case of Maximum Selectiveness: We first look at the cases of maximum selectiveness. The first two sequences contain each one goal while no goals were scored in the sequence in Fig. 7. Just like for the test sequence in Fig. 4, the peaks in the highlights time curves in Figs. 5(b) and 6(b) indicate the goals as the only events in these sequences satisfying the high requirements on the excitement "strength" posed by setting $M = 3$. On the other hand, no such peaks were found in the highlights time curve in Fig. 7(b). This consistence in the composition of highlighting video abstracts generated in a uniform way for video sequences that were taken under different conditions (different broadcasters) and where the goals were scored in different game contexts, can be seen as a first important indication of the robustness of the proposed approach. Namely, although only four goals were found in the test sequences, they differed strongly from each other, both in terms of realization and coverage. For instance, the first goal in Fig. 4 and the goal in Fig. 6 resulted from the actions of a similar type (massive attack toward the goal emerged from the normal course of the game). However, their realizations differed in the sense that the first action came from a side and the other one from the center of the field. As opposed to these goals, the goal in Fig. 5 resulted from a corner kick. The coverage of this goal was different from other goals as in addition to the standard wide-range and close-up shots of the players on the field, shots were also taken by the camera positioned behind the goalkeeper and by another one that focused on the player performing the corner kick. Finally, the second goal in Fig. 4 was scored from a quick action that emerged from a game break. The presence of the game break and a specific course of action toward the goal resulted in a coverage that was different than the coverage of other goals in our test set.

An additional indication of the robustness of the proposed approach for the case of maximum selectiveness can be drawn from the fact that the peaks indicating the goals in Figs. 4(b), 5(b), and 6(b) are all in the value range of medium-to-high

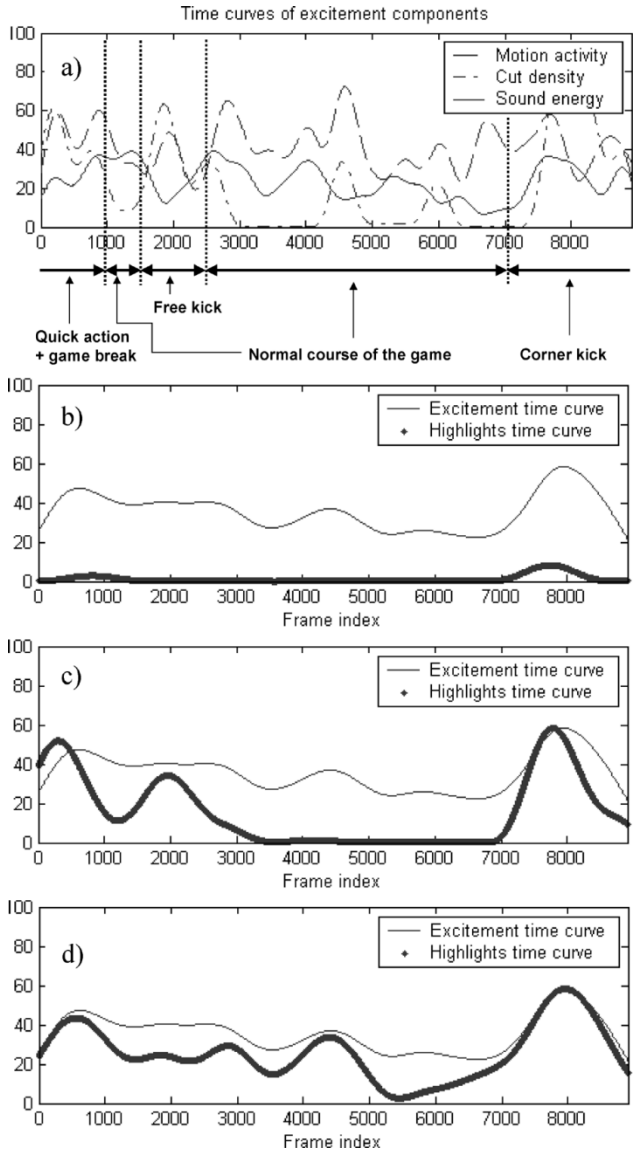

Fig. 7. (a) Component time curves obtained for an excerpt from a soccer match. Excitement and highlights time curve obtained for the case of (b) $M = 3$, (c) $M = 2$, and (d) $M = 1$.

excitement (40%–80% on the excitement value scale) and thus clearly detectable, in spite of the differences in broadcasters and in the goal realization and coverage. Constructing comparable highlights time curves of the sequences taken under different conditions is enabled by the scaling and normalization mechanisms introduced in the definitions of the component time curves [10]. In addition to enabling the component time curves of different modalities to be compared to each other and to be combined into the excitement and highlights time curves, these mechanisms also reduce the sensitivity of the curves with respect to the varying technical parameters, such as the basic sound volume level or frame resolution. The unavoidable residual variations of excitement values at the goal-peaks are mainly due to the factors such as the directing style (e.g., long versus short shots, close-up versus wide-range shots, fixed versus moving camera), different styles of the commentators (temperament) and some other technical parameters (e.g. camera position and dynamics). While these variations can be explained in part by different conventions of coverage adopted by broadcasters, another cause of these variations we see in the differences in the realizations of one and the same event. If we take the "goal" event as an example, a goal can be scored, for instance, after a surprise attack, after a corner kick or a free kick. Depending on the dynamics of an action and the presence and length of a game break preceding or succeeding the goal, these different realizations of the "goal" event will be characterized by different cheering and motion activity patterns, different directing (editing) practices and different degrees of their temporal synchronization. This directly translates to the behavior of the component time curves and the resulting highlights time curves where the peaks of similar events will have (slightly) different heights.

The differences in coverage of various event realizations can also be related to the varying importance of these realizations. As namely not every goal is equally important, the more important goals (like those determining which team will enter the next round of a championship) will most likely be characterized by the commentator and audience being more loud, and by a longer period of celebration which is, in most cases, covered by a longer sequence of short and dynamic close-up shots. This will result in higher peaks of the component and highlights time curves at these video segments and thus indicate the periods of "strong" excitement as compared to a lower level of excitement obtained for some other, less important goals.

Based on the discussions above, we may conclude that a highlights time curve is not necessarily equivalent to a specific event detector. In our soccer case study the events such as a spectacular action or free kick can be characterized by stronger excitement than some goals. Also, the excitement strength of some unimportant goals may not be sufficient for their inclusion in the highlighting video abstract in the case of maximum selectiveness. Then, these spectacular nongoal events may be made extractable for $M = 3$ while the (unimportant) goals may be made accessible only after lowering the selectiveness criterion. The highlights time curve generated for soccer and for the case of maximum selectiveness should therefore not be seen simply as a "goal detector", but more generally, as a detector of a subset of events (which ever these may be) that are characterized by sufficiently strong excitement. The generic nature of a highlights time curve will become more visible when we analyze the cases of less selectiveness in the next section.

We now analyze the sequence considered in Fig. 7 to investigate whether the flatness of the highlights time curve for $M = 3$ can be justified by the absence of sufficiently exciting events. As can be seen from Fig. 7(b), this curve is almost flat, with the highest value being less than 10% on the excitement value scale. An analysis of this sequence and its comparison with other test sequences showed that the attack toward the goal in the beginning of this sequence was of a similar type as the events around the frame 10 000 in Figs. 4 and 5 and the one at the beginning of the sequence in Fig. 6: they were all characterized by quick attacks that were stopped either by a foul play or the ball going out of the field. Further, the free kick in Fig. 7 was similar to the free kick in Fig. 4, both regarding the situation on the field and the coverage. Finally, the corner kick at the end of the sequence was covered in a similar way as the one resulting in a goal in

Fig. 5. That this corner kick was not successful and therefore also less exciting than the one in Fig. 5 is properly indicated by the level of sound energy which is considerably lower than in Fig. 5 and which prevents this event to be included into the events of the same class as the successful corner kick in Fig. 5.

2) Less Selective Cases: If we now look at the changes in highlights time curves in Figs. 5–7 as a consequence of the changing value of M, we see similar effects as in Fig. 4. The scope of extractable events increases, and the ratio of the extracted material per event changes for different values of M. As we could see from the content labels in the Figs. 5(a), 6(a), and 7(a), the filtering process works in most cases according to expectations. For instance, all peaks to be found starting from the frame 4000 in Fig. 5(c) and (d) correspond to distinct events that are more and more "amplified" with the reduction of the value of M. The small peak around the frame 1000 in Fig. 5(c) and its amplified version in Fig. 5(d) indicate no specific event, but a particularly dynamic game segment with several good actions following each other and covered by zoom-ins on players and duels on the field. Similar type of dynamic ball exchange also characterizes the segment of the sequence in Fig. 7 between frames 3000 and 5000, which is indicated by medium-range excitement values in Fig. 7(d). An intriguing aspect of Fig. 5 is the rather low peak around the frame 14 000 in Fig. 5(d). This peak correctly indicates a free kick and is lower than the peaks indicating the realizations of the "free kick" event in the sequences in Figs. 4 and 7. We explain this difference by the fact that the free kick in Fig. 5 was performed after a relatively long game break that was caused by a controversial foul play. The game break was covered by a series of dynamic close-up shots of moving players which is indicated by the peaks in the highlights time curves in Fig. 5(c) and (d) around frame 13 000. After this break, the free kick is performed rather quickly and the game is immediately resumed. In this sense, we may also say that here the foul play and the resulting game break were more interesting than the free kick itself, which is correctly indicated by the highlights time curves.

In Fig. 6(c), the quick action in the first segment of the sequence is correctly indicated as a highlight candidate for $M = 2$. This is, besides the goal, the only distinct event to be considered as a highlight candidate in the first half of the test sequence. As we already mentioned in the previous section, this action is of a similar type as the actions also found in Figs. 4, 5, and 7. Although these four realizations of the "quick action" event took place in different sectors of the field and were covered in different ways in terms of camera activity, level of zoom and the visual content captured, they were treated in a similar way by our approach. This can be seen in the similar behavior of the highlights time curves in Fig. 4(c) and (d) (around frame 10 000), Fig. 5(c) and (d) (around frame 10 000), Fig. 6(c) and (d) (first 1000 frames), and Fig. 7(c) and (d) (first 1000 frames) in terms of the reached excitement levels which are all in the range of medium excitement (40%–60% on the excitement scale). In the second half of the sequence in Fig. 6, the small peak around the frame 13 000 in Fig. 6(b) indicates an interesting fast action toward the goal that finishes by a nice move of the goalkeeper preventing the score. Although it is (correctly) not considered the strongest highlight, one would expect this peak to be amplified for lower values of M. This, however, is not the case. Moreover, this event is surrounded in Fig. 6(c) by other two peaks indicating the player exchange and the replay of the action mentioned above, respectively. This "imperfection" was caused by the high rhythm and motion activity components in the surrounding segments which outperformed the sound energy increase caused by cheering the action in the actual action segment. However, due to smoothing, the entire sequence of these three events is then recognized in Fig. 6(d) as one long highlight candidate, which is probably the best solution in view of a rather dynamic character of this second part of the sequence. Only the parts around the frames 3000 and 6000 were left out in all cases. This is acceptable as these parts correspond to stationary game segments.

V. Conclusions

With the method for highlights extraction from a sport video that we proposed in this paper, we intended to come one step closer to a generic solution of the highlights extraction problem. The generic aspect of the proposed method is best visible from its underlying concept of "excitement": instead of modeling each potential highlighting event separately, we simply search for the video segments where the excitement evoked in the user is expected to be sufficiently high.

The proposed approach consists of five major steps: 1) feature extraction; 2) computation of the component time curves; 3) adaptive filtering; 4) highlights time curve generation; and 5) applying a cutoff line to extract highlights in a prespecified length. The first two steps should, in general, be considered specific for each particular sport program genre. However, once this "raw information" is available, the fully generic highlight extraction steps (3–5) can be applied. Regarding step 1), it is worthwhile emphasizing that the feature set used in this paper, although generated carefully, may still be suboptimal. We see the problem not that much in the selection of features that are already in the set but mainly in the small size of this set. It is namely so that our combined analysis of features proves to be beneficial for reducing the imperfections of single features. Indeed, the results obtained in our "soccer" case study are not bad, and more importantly, they are encouraging. However, the robustness of the approach may increase with more evidence extracted from the audiovisual data stream. To extend our feature set, new findings are needed regarding the links between the audio-visual stimuli and human affective responses. Establishing these links is a subject of current and future research in the field of psychophysiology.

References

[1] Durlacher Research Ltd., "Digital Local Storage: PVR's, Home Media Servers, and the Future of Broadcasting," report, Nov. 2000.

[2] B. Adams, C. Dorai, and S. Venkatesh, "Toward automatic extraction of expressive elements from motion pictures: tempo," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 472–481, Dec. 2002.

[3] N. Babaguchi and N. Nitta, "Intermodal collaboration: A strategy for semantic content analysis for broadcasted sports video," in *Proc. IEEE ICIP 2003*, Barcelona, Spain, Sep. 2003.

[4] Y.-L. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated image and speech analysis for content-based video indexing," in *Proc. 3rd IEEE Int. Conf. Multimedia Computing and Systems*, 1996, pp. 306–313.

[5] S. Dagtas and M. Abdel-Mottaleb, "Extraction of TV highlights using multimedia features," in *Proc. 4th IEEE Workshop on Multimedia Signal Processing*, 2001, pp. 91–96.
[6] B. H. Detenber, R. F. Simons, and G. G. Bennett, "Roll 'em!: the effects of picture motion on emotional responses," *J. Broadcast. and Electron. Media*, vol. 21, pp. 112–126, 1997.
[7] A. Ekin and A. M. Tekalp, "Robust dominant color region detection and color-based applications for sports video," in *Proc. IEEE ICIP 2003*, Barcelona, Spain, Sep. 2003.
[8] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. ICMCS '95*, May 1995, pp. 167–174.
[9] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *Proc. IEEE Int. Conf. Multimedia and Expo 2002 (ICME '02)*, vol. 1, Aug. 26–29, 2002, pp. 821–824.
[10] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, no. 1, pp. 143–154, Feb. 2005.
[11] A. Jaimes, T. Echigo, M. Teraguchi, and F. Satoh, "Learning personalized video highlights from detailed MPEG-7 metadata," in *Proc. IEEE ICIP 2002*, vol. 1, Rochester, NY, 2002.
[12] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," in *Proc. Int. Conf. Image Processing 1998 (ICIP 98)*, vol. 1, Oct. 4–7, 1998, pp. 871–874.
[13] R. Leonardi, P. Megliorati, and M. Prandini, "Semantic indexing of sports program sequences by audio-visual analysis," in *Proc. IEEE ICIP 2003*, Barcelona, Spain, Sep. 2003.
[14] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001)*, 2001, pp. 132–138.
[15] ——, "Semantic sports video analysis: approaches and new applications," in *Proc. IEEE ICIP 2003*, Barcelona, Spain, Sep. 2003.
[16] N. Nitta, N. Babaguchi, and T. Kitahashi, "Extracting actors, actions and events from sports video—a fundamental approach to story tracking," in *Proc. 15th Int. Conf. Pattern Recognition (ICPR 2000)*, vol. 4, pp. 718–721.
[17] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE ICASSP*, Salt Lake City, UT, May 2001.
[18] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multimodal extraction of highlights from TV formula 1 programs," in *Proc. IEEE Int. Conf. Multimedia and Expo 2002 (ICME '02)*, vol. 1, pp. 817–820.
[19] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
[20] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia 2000*, Los Angeles, CA, 2000.
[21] C. G. M. Snoek and M. Worring, "Time interval maximum entropy based event indexing in soccer video," in *Proc. IEEE Int. Conf. Multimedia and Expo 2003 (ICME '03)*, vol. 3, pp. 481–484.
[22] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. IEEE Int. Workshop on Content-Based Access of Image and Video Database 1998*, pp. 81–90.
[23] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka, "An object detection method for describing soccer games from video," in *Proc. 2002 IEEE Int. Conf. Multimedia and Expo 2002 (ICME '02)*, vol. 1, Aug. 26–29, 2002, pp. 45–48.
[24] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing 2002 (ICASSP '02)*, vol. 4, pp. IV-4096–IV-4099.
[25] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2003, pp. 632–635.
[26] G. Xu, Y.-F. Ma, H.-J. Zhang, and S. Yang, "A HMM based semantic analysis framework for sports game event detection," in *Proc. IEEE ICIP 2003*, Barcelona, Spain, Sep. 2003.

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

He is an Associate Professor and Head of the Multimedia Content Analysis research cluster at the Department of Mediamatics, Delft University of Technology. He was a Research Fellow at British Telecom Labs, Ipswich, U.K., in 2000–2001, and a Visiting Scientist at Hewlett-Packard Labs, Palo Alto, CA, in 1998, at Philips Research Labs, Briarcliff Manor, NY, in 2003, and at Microsoft Research Asia, Beijing, China, in 2005. His research interests and expertise are in the broad areas of multimedia signal processing, multimedia computing, 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 has authored and co-authored more than 50 publications, including the books titled *Image and Video Databases: Restoration, Watermarking and Retrieval* (Elsevier, 2000) and *Content-Based Analysis of Digital Video* (Kluwer, 2004).

Dr. Hanjalic was a Guest Editor of the *International Journal of Image and Graphics*, Special Issue on Content-based Image and Video Retrieval, July 2001, and has served as a Program Committee member, session chair, or a panel member in many international conferences and workshops, such as IEEE ICME, IEEE ICIP, IEEE ICASSP, ICCV, ACM Multimedia, ACM SIGIR, ACM SIGMM MIR, CIVR, IS&T/SPIE Storage and Retrieval for Media Databases, and IS&T/SPIE Internet Multimedia Management Systems. He was the initiator and main organizer of the Symposium on Multimedia Retrieval, Eindhoven, The Netherlands, in January 2002. He is a Dutch representative in the Management Committee and a workgroup leader of the EU COST 292 action "Semantic Multimodal Analysis of Digital Media". He also serves regularly as an advisor/reviewer of the Belgian Science Foundation (IWT) for project proposals in the area of Information Technology and Systems. From 2002 to 2005, he was the secretary of the IEEE Benelux Section. He was also a member of the Organizing Committee of the IEEE International Conference on Multimedia and EXPO (ICME) 2005, and will be a Co-Chair of the new IS&T/SPIE Conference on Multimedia Content Analysis, Management and Retrieval 2006, San Jose, CA, January 2006.