AUTOMATIC SUMMARIZATION OF AUDIO-VISUAL SOCCER FEEDS

*Fan Chen, C. De Vleeschouwer**

Université catholique de Louvain
ICTEAM
Louvain-la-Neuve, Belgium
Email: christophe.devleeschouwer@uclouvain.be

H. Duxans Barrobés, J. Gregorio Escalada, D. Conejero

Telefónica Investigación
y Desarrollo
Barcelona, Spain
Emails: {hdb, jges, dco}@tid.es

ABSTRACT

This paper presents a fully automatic system for soccer game summarization. The system takes audio-visual content as an input, and builds on the integration of two independent but complementary contributions (i) to identify crucial periods of the soccer game in a fully automatic way, and (ii) to summarize the soccer game as a function of individual narrative preferences of the user. The process involves both audio and video analysis, and handles the personalized summarization challenge as a resource allocation problem. Experiments on real-life broadcasted content demonstrate the relevance and the computational efficiency of our integrated approach.

***Keywords*—** Audio & video analysis, summarization.

1. INTRODUCTION AND OVERVIEW

This paper considers sport event summarization. The purpose is the generation of a concise video with well-organized and personalized story-telling. Although many works have been devoted to the automatic detection of key actions in football games [1, 2, 3], little attention has been given to the construction of a summary telling a story and including all events that satisfy individual user interest. Actually, when addressing the problem of building a summary from highlighted actions, most earlier methods focus on events retrieval scenarios, and just implement pre-defined filtering or ranking procedures to extract the actions of interest (or the ones that match the theme requested by the user) from the original audiovisual stream. Most methods are definitely rigid in the sense that the pre-encoded summarization scheme can not be adapted to any kind of user preferences, neither in terms of preferred action, nor in terms of the desired length and narrative style of the summary. Some of them, e.g. [3], order segments in decreasing order of importance, and can thus easily handle distinct summary length constraints. However, it just arbitrarily extracts a pre-defined fraction of the scene (e.g. 15 or 30 seconds prior the end of the last live action segment preceding the replay), without taking care of story-telling artifacts.

Our work attempts to overcome those limitations by introducing a generic resource allocation framework to adapt the selection of audio-visual segments to be included in the summary according to the needs and preferences of the user. Several contending local stories are considered to present each segment, so that not only the content, but also the narrative style of the summary can be adapted to user requirements. Hence, by tuning the benefit and the cost of the local stories, our framework becomes able to balance -in a natural and personal way- the semantic (what is included in the summary) and narrative (how it is presented to the user) aspects of the summary. This is a fundamental difference, compared to the approaches based on filtering or ranking mechanisms. Interestingly, and in contrast to [4], our framework also avoids managing sophisticated semantic relations among concept entities to control story-telling. This enables the deployment of an effective system even when limited semantic understanding of the content is available, e.g. in cases for which relevant periods of the game are detected based on broadcasted audio analysis.

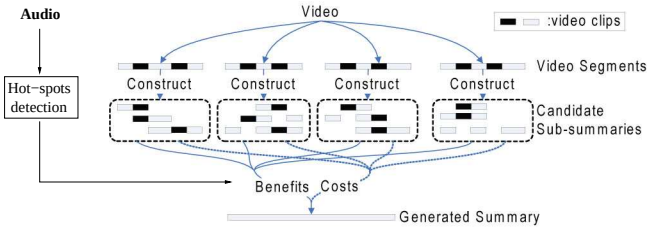


Fig. 1. System overview. On the left, audio feed is analyzed to detect highlighted events. On the right, video clips detection and corresponding view types recognition supports segmentation and local stories organization. On the bottom, summarization is implemented as a resource allocation problem.

Our proposed system is depicted in Fig.1. It consists in three main steps. First, in Section 2, the instantaneous energy and the local periodicity of the audio signal is measured to infer highlighted moments in the game. Second, in Section 3, the video is split into clips, based on conventional shot (or clip) boundary detection tools. The view type structure of the clips composing the video signal is then investigated to divide the audio-visual feed into non-overlapping and semantically meaningful segments, i.e. into segments which are coherent with the actions of the game. Hence, we segment the video by analyzing patterns of camera switching and view types, i.e. based on the monitoring of production actions, rather than based on (complex) semantic scene analysis tools. As a third and last step, in Section 4, our system envisions personalized summarization as a resource allocation problem. Specifically, if we define a sub-summary or local story as one way to select clips within a segment, we can regard the final summary as a collection of non-overlapping sub-summaries. Knowing view types of clips and highlighted moments in the game, Section 4 considers all possible combinations of clips for each segment, and describes how to account for individual narrative preferences to assign benefits to each one of those local stories. Summarization is then implemented as the search for the combination of sub-summaries that maximizes the overall (user-dependent) benefit under a global duration constraint. Results and conclusions are presented in Sections 5 and 6, respectively.

*UCL's contribution has been partly funded by the European FP7 APIDIS project, the Walloon Region WALCOMO project, and the Belgian NSF.

2. HIGHLIGHTED MOMENTS IN AUDIO FEED

A preliminary step in generating a soccer game summary consists in detecting special events, also called hot spots. Those events include all actions that induce a reaction from the commentator or the public, e.g. a goal or a spectacular dribbling.

This section explains how the soccer game hot spots can be detected based on the analysis of the broadcasted audio signal (including both the ambient noise and the commentator voice). Compared to earlier contributions [5, 6], our system has the advantage (i) to process only the audio signal, thereby resulting in a computationally efficient system, (ii) to exploit both the audio commentary and the ambient noise, through language independent features, and (iii) to avoid the need for training material or manual tuning.

Since a detailed motivation of the method has been provided in one of our earlier publication [7], we only remind its key principles and formal definition for completeness.

The location and significance of each hot spot is computed based on two acoustic features, directly extracted from the audio track, on a set of consecutive and partly overlapping windows. The first feature, denoted block energy, reflects the squared sum of the signal contained in the window of interest, between 660Hz and 4400KHz. The second one, denoted repetition index, measures how well the central audio pattern is repeated along the observation window, through cross-correlation computations. High values of repetition index typically reflect short words (e.g. "goal" or the name/nickname players) repetitions by the commentators, or lengthening of vowels, for example the "o" in the word "goal", after a significant event.

Formally, the block energy and the repetition index are computed for each second of the game. Each block energy value results from the median filtering of size 3 of the maximal squared sum value measured on a 4-uple of 75% overlapping one-second-length windows. To compute the repetition index, the audio track is divided in 2 seconds length windows, with a 50% of overlap. The central part of each window is selected as the narrow acoustic section of interest, and the normalized cross correlation between this section and the rest of the window is computed. Finally, the repetition index is defined to be the mean of the 5 highest values of the cross-correlation.

Once the acoustic features have been computed, highlighted events and corresponding significance values have to be defined. Highlighted events are first extracted through an iterative and greedy process. At each step, the one-second interval with highest block energy defines an additional highlighted event, and a masking window prevents the future selection of other hot spots, related to the same game event, in a 20 seconds neighborhood around the highlighted event.

The significance of each highlighted event is then computed as a function of both the block energy and repetition index. For those instants for which the repetition index is higher than the threshold $Th_{RepIndex}$, the score is simply set to the energy block value. In contrast, for those instants for which the repetition index is lower than the threshold $Th_{RepIndex}$, the score is decreased to 95% of the energy block value. The threshold $Th_{RepIndex}$ is determined based on the statistical distribution of the repetition index, i.e. $Th_{RepIndex} = \mu_{RepIndex} + 1.5\sigma_{RepIndex}$, with $\mu_{RepIndex}$ and $\sigma_{RepIndex}$ denoting the mean and standard deviation of the repetition index, respectively.

3. VIDEO SEGMENTATION

We define a video segment as a period covering several successive clips closely related in terms of story semantic, e.g., clips for an attacking action in football including both a free-kick and its consequence. By considering construction of sub-summaries in each segment independently, we trade-off summarization between efficiency of computation and controllability of story organization. In Fig.2, we explain the segmentation rule that we envision for sport videos. It is worth noting that the proposed segmentation do not assume any complex hand-made annotation process, or sophisticated automatic analysis of the video sequence. Rather, we exploit the fact that the transitions in the state of the game motivate camera or view type switching, and are thus reflected in the production actions[8]. Hence, similarly to the approach presented in one of our earlier publication [9], we segment the video based on the monitoring of the production actions, instead of (complex) semantic scene analysis tools.

A general diagram of state transition in one round of offense/defense is given in Fig.2, together with the view type structure of the corresponding video. It optionally starts with a kick-off type of action. In that case, a close view is used to highlight the player who kicks off. Then, the offensive side makes trials of score after several passing actions, rendered through far or medium views. This trial ends up in one of three possible results: being intercepted, scoring, or being an opportunity.

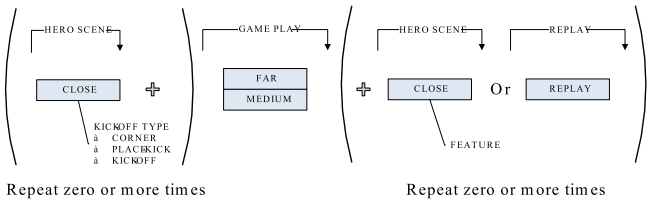


Fig. 2. Soccer game video segmentation is based on view-type subsequence matching.

After a key event is finished, some close-up shots might be given to raise the emotional level. According to the importance of the corresponding event, replay clips might also be appended. Note that close view, medium view and replay are all optional. We regard the state chain from the action-start to one of the results as a semantically complete segment, and divide the video into a series of segments based on observation of view types and recognition of kick-off actions. Although there are some complex cases, e.g., the offensive side tries many times of shooting, our segmentation rule is still applicable to them, because the producer will not switch the view type during those periods due to the tightness of match. A similar analysis of the video in view-types was used in [3] to help detect exciting events(i.e., game parts with both close-up scenes and replays).

4. RESOURCE-CONSTRAINED SUMMARIZATION

We investigate summarization as a resource allocation problem. It first involves the definition of several contending strategies to render each individual video segment (Section 4.1). It then implies the optimal allocation of a global summary duration among the available strategies (Section 4.2).

4.1. Sub-summaries definition

We now explain the construction of sub-summaries by clip selection within a segment. The purpose of this section is two-fold. First, we explain how to define distinct sub-summaries (also named local stories) based on the knowledge of view type and scene type for each one of the segment clips. Second, we derive a benefit and a cost metric for each sub-summary, to be used during resource allocation.

In the following, we consider the m^{th} segment, and explain how its sub-summaries, and associated costs and benefits, are computed. For notation simplicity, we omit the index m of the segment under

investigation. Mathematically, assume that the segment is composed of N consecutive clips. For the i-th clip, we identify its scene type s_i, view type v_i, and replay status r_i. We set $s_i = 0$ for public scene and $s_i = 1$ for game scene, and set v_i to $0, 1, 2$ for the close, medium and far view, respectively. $r_i = 0$ for the i-th clip in the replay mode and $r_i = 1$ for normal game. We also assume that a level of interest $\mathcal{I}_i$ has been computed for each clip i, based on the propagation of the audio significance to individual clips, as described later in this section. We then use $a_{ki} = 1$ to represent the adoption status of selecting the i-th clip into the k^{th} sub-summary of the segment, and $a_{ki} = 0$ for not selecting the clip. The cost of $\mathbf{a}_k$ is its length, which is defined as $|\mathbf{a}_k| = \sum_{i|a_{ki}=1} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|$. We use τ_i^{S} and τ_i^{E} to denote the starting time and ending time of the i-th clip in the broadcasted video. We then propose to define the benefit gained by using $\mathbf{a}_k$ to define the k^{th} sub-summary of the segment as follows

$$\mathcal{B}(\mathbf{a}_k) = \sum_{i|a_{ki}=1} \mathcal{I}_i \left(1 + \phi(a_{k(i-1)} + a_{k(i+1)})\right) \mathcal{P}(\mathbf{a}_k). \quad (1)$$

The term of $a_{k(i-1)} + a_{k(i+1)}$ is included to add extra benefit from continuity of story-telling by selecting pairs of successive clips. Parameter ϕ controls the importance of story continuity, where a higher ϕ leads to more continuous summaries. $\mathcal{P}(\mathbf{a}_k)$ represents the penalty brought by redundancy in the sub-summary and other forbidden cases, e.g., the replay is selected while no normal play part is selected, i.e.,

$$\mathcal{P}(\mathbf{a}_k) = \left[(1-\alpha) \frac{\sum_{i|a_{ki}=1, v_i>0, r_i=1} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|}{\sum_{i|a_{ki}=1} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}|} + \alpha \right]^{\gamma} \mathcal{P}_2(\mathbf{a}_k). \quad (2)$$

where the term in the bracket shows the rate of redundancy brought by replays and close views. Remember here that the i-th clip is NOT a replay if $r_i = 1$, and is NOT a close view if $v_i > 0$. Hence, higher γ tolerates less redundancy, which produces a summary with more but shorter sub-summaries with less replays, while lower γ produces a summary including less but longer sub-summaries with detailed replays. $\mathcal{P}_2(\mathbf{a}_k)$ switches its value between 0.1 and 1, according to whether $\mathbf{a}_k$ is of a forbidden form or not. In our following implementation for soccer videos, we have defined the following forbidden cases. 1) A sub-summary with only replays. 2) Kick-off action without the first far/medium view clip. 3) From the ending time of the final far/medium view in the segment, the continuous period in the summary for explaining the consequence of the action is shorter than a given length (5 seconds in our experiments).

Those heuristic definitions are motivated by general production principles, which promote continuous story-telling and trade-off local and global completeness. It is worth mentioning here that searching for an optimal benefit function is probably utopic. Instead, we believe that any function that is able to capture and trade-off the subjective notions of redundancy and completeness, while promoting continuous story-telling, provides a valid alternative to the formulation provided in equation (1).

To complete this section, we now explain how the significance associated to the audio-highlighted moments detected in a segment can be translated into an interest $\mathcal{I}_i$ for the i^{th} clip of the segment. Again, heuristic and good-sense strategies are proposed.

To account for heterogeneous user preferences regarding the inclusion of highly emotional clips, i.e. close views and replays, we derive both a game relevance $\mathcal{I}_i^G$ and an emotional level $\mathcal{I}_i^E$ for the i-th clip. The interest level $\mathcal{I}_i$ is then computed as

$$\mathcal{I}_i = \alpha \mathcal{I}_i^E + (1-\alpha)\mathcal{I}_i^G; \quad (3)$$

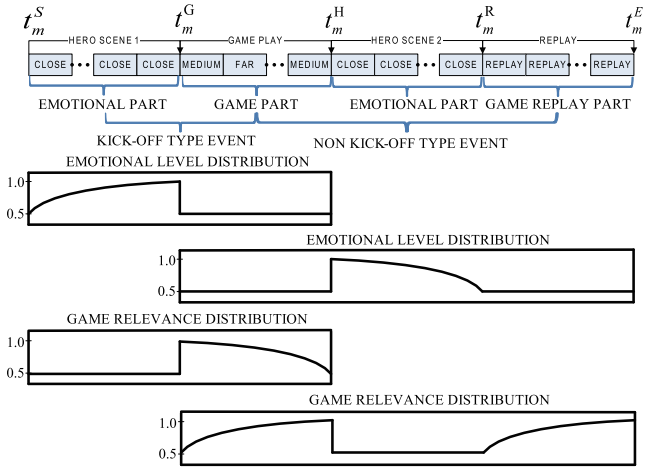


Fig. 3. Game relevance and emotional level are assigned as a function of clips view-types.

α is a hyper-parameter of user preference to control the relative importance of emotional level and game relevance.

Considering the production principles, we define $\mathcal{I}_i^G$ and $\mathcal{I}_i^E$ by propagating the significance of a particular moment according to the view type structure of the segment, as depicted in Fig.3 for the m^{th} segment.

Formally, adopting the above notations regarding the status of a clip, and assuming that we have L events annotated for the m^{th} segment, we design the distribution of game relevance and emotional level within the segment as shown in Fig.3, where the percentage of game relevance $\mathcal{D}_{li}^G$ and emotional level $\mathcal{D}_{li}^E$ induced by the l-th event on the i-th clip satisfy $1 = \sum_i \mathcal{D}_{li}^E = \sum_i \mathcal{D}_{li}^G$, and read

$$\mathcal{D}_{li}^E \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + \delta_{v_i,0} \tanh\left(\beta \frac{t - t_m^S}{t_m^G - t_m^S}\right) dt, \quad (4)$$

$$\mathcal{D}_{li}^G \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + (1 - \delta_{v_i,0}) \tanh\left(\beta \frac{t_m^H - t}{t_m^H - t_m^G}\right) dt \quad (5)$$

for kick-off events, and

$$\mathcal{D}_{li}^E \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + \delta_{v_i,0} \tanh\left(\beta \frac{t_m^R - t}{t_m^R - t_m^H}\right) dt, \quad (6)$$

$$\mathcal{D}_{li}^G \propto \int_{\tau_i^{\mathrm{S}}}^{\tau_i^{\mathrm{E}}} 1 + (1 - \delta_{v_i,0})\delta_{r_i,1} \tanh\left(\beta \frac{t - t_m^G}{t_m^H - t_m^G}\right) + (1 - \delta_{v_i,0})\delta_{r_i,0} \tanh\left(\beta \frac{t - t_m^R}{t_m^E - t_m^R}\right) dt \quad (7)$$

for non-kick-off events, where t_m^G, t_m^H, t_m^R are starting times of game play part, hero scene part, and replay part of the m-th segment, respectively, as shown in Fig.3. β is a parameter introduced to control the decaying speed and $\delta_{x,y}$ is the Kronecker delta. We use hyperbolic tangent function $\tanh(x)$ to model the decaying process, because it is bounded and is integrable which simplifies the computation of $\mathcal{D}_{li}^E$ and $\mathcal{D}_{li}^G$.

We motivate the above heuristic definition by considering production principles for team-sports. For a kick-off type of event, which takes place in the beginning of the segment, we limit its influence within the first hero-scene and the game play part. Since the dominant player appears at the end of hero scene and commits his action in the beginning of game play part, it is natural to let emotional level increase along with time evolving in the close view, and let game relevance decrease in the game play. For non-kick-off game events, we design their distributions in Fig.3 based on the following facts: 1) Close views and replays are appended right after a critical action. A clip closer to the hero scene or replay has a higher game relevance. 2) The hero scene usually starts with the most highlighted player and then moves to other less important objects. Accordingly, the emotional level in the hero scene should decrease along with time evolving. 3) The replay part is designed to play the critical action in the same temporal order, which means that the game relevance is also increasing along with the evolving of replay. For other events, such as public events, player exchange, and medical assistance, we assign uniform game relevance over its game part and uniform emotional level over its close-view part.

Game relevance $\mathcal{I}_i^G$ and emotional level $\mathcal{I}_i^E$ of the i-th clip are then computed by accumulating all related events, i.e.,

$$\mathcal{I}_i^E = \mathcal{T}_m^E \sum_l \mathcal{D}_{li}^E \mathcal{G}_l, \qquad \mathcal{I}_i^G = \mathcal{T}_m^G \sum_l \mathcal{D}_{li}^G \mathcal{G}_l, \tag{8}$$

where $\mathcal{G}_l$ represent the level of significance assigned to the l-th audio-highlighted event. $\mathcal{T}_m^G$ is the length of game play in the segment, which includes all far/medium views, i.e., $\mathcal{T}_m^G = \sum_{i,v_i>0} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}| s_i$. $\mathcal{T}_m^E$ is the length of emotional highlights, which consists of all non-replay close views, i.e., $\mathcal{T}_m^E = \sum_{i,v_i=0} |\tau_i^{\mathrm{E}} - \tau_i^{\mathrm{S}}| s_i r_i$. $\mathcal{T}_m^G$ and $\mathcal{T}_m^E$ are introduced to avoid to favor short actions too much during the resource allocation process. [1]

4.2. SUB-SUMMARIES ALLOCATION

We now explain how the global summary duration resource is allocated among the available local sub-summaries to maximize the aggregated benefit. Formally, assume that a video is divided into M segments. Collecting at most one sub-summary for each segment, we form the final story. Here, in contrast to Section 4.1, we refer explicitly to the index m of the segment, and let $\mathbf{a}_{mk}$ denote the k^{th} sub-summary of the m-th segment. The overall benefit of the whole summary is defined as accumulated benefits of all selected sub-summaries, i.e.,

$$\mathcal{B}(\{\mathbf{a}_{mk}\}) = \sum_m \mathcal{B}(\mathbf{a}_{mk}) \tag{9}$$

with $\mathcal{B}(\mathbf{a}_{mk})$ being defined in equation (1) as a function of the user preferences, and of the highlighted moments. Our major task is to search for the set of sub-summaries indices $\{k^*\}$ that maximizes the total payoff $\mathcal{B}(\{\mathbf{a}_{mk}\})$ under the length constraint $\sum_m |\mathbf{a}_{mk}| \leqslant u^{\mathrm{LEN}}$, with u^{LEN} being the user-preferred length of the summary.

This resource allocation problem has been extensively studied in the literature, especially in the context of rate-distortion (RD) allocation of a bit budget across a set of image blocks characterized by a discrete set of RD trade-offs [10, 11]. Under strict constraints, the problem is hard and relies on heuristic methods or dynamic programming approaches to be solved. In contrast, when some relaxation of the constraint is allowed, Lagrangian optimization and convex-hull approximation can be considered to split the global optimization problem in a set of simple block-based local decision problems [10, 11]. The convex-hull approximation consists in restricting the eligible summarization options for each sub-summary to the (benefit,cost) points sustaining the upper convex hull of the available (benefit, cost) pairs of the segment. Global optimization at the video level is then obtained by allocating the available duration among the individual segment convex-hulls, in decreasing order of benefit increase per unit of length. This results in a computationally efficient solution that can still consider a set of candidate sub-summaries with various descriptive levels for each segment.

5. EXPERIMENTS AND RESULTS

The objective and quantitative results presented herebelow only refer to a single soccer game. This is because the manual creation of accurate reference annotations is a time consuming process, which has only been carried out for this particular game. During the conference however, we plan to present several short summaries that have been generated automatically from 4 soccer videos, initially broadcasted according to UEFA rules.

As a first step of our proposed summarization procedure, the input video is cut into segments, based on clips view-type analysis, as described in Section 3. The clips are defined based on a standard shot-boundary detection method [12]. View-types are then defined based on the method in [1], and replays are detected based on specific logo appearance. Due to size limitation of supplement file, we cannot provide the segments extracted from this process in the zip file[13]. They will however be presented on the web page associated to this paper. Without surprise, we observe that, in absence of close views, the segmentation ends up in merging several rounds of offense/defense actions in a single segment. However, we observe in the summary examples[13] that it does not penalize the summarization process. This is because all critical actions end with close views, so that the propagation of audio significance effectively propagates the clips that end important actions and are close to audio-highlighted instants.

To assess the relevance of the proposed audio analysis process, we have compared the fifties most significant audio detected hot spots with a manually generated annotation. For each video segment, the manual annotation identifies all occurrence(s) of actions within a predefined set of soccer game actions (as defined on top of Fig.4). We are interested in measuring which ones of those reference and manually defined actions are actually identified by the automatic audio hot spot detection module. For this purpose, Fig.4 presents the recall rate and the frequency of occurrence for each type of action. In addition, it presents the average level of significance of audio detected actions. We observe that the goals have the highest level of significance, and that all actions have reasonable recall rates, which is due to the fact that audio intensity does not strictly focus on one type of action, but rather identifies tight and relevant periods of the game.

We now assess the behavior of our proposed resource allocation summarization framework.

In Figure 5, we compare the two benefit-cost convex-hulls obtained with two distinct α parameters, for the same segment of the game. At the top of each individual figure, we first plot the set of (benefit,cost) pairs for all possible local stories of the segment[2], for a given α parameter. In each figure, the so-called 'convex-hull optimal' sub-summaries are the ones lying on the upper convex-hull of all (benefit, cost) points, and are denoted with circles. In contrast, the vertical blue lines represent other sub-optimal local stories. We observe that only a few combination of clips, i.e. a few candidate sub-summaries, lie on the convex-hull. Hence, despite there are many different ways to select the clips within the segment, our framework naturally ends up in selecting a small subset of eligible and convex-hull optimal sub-summaries,

[1] Replacing $\mathcal{T}_m^G$ and $\mathcal{T}_m^E$ in Eq.8 would enable extra controllability regarding the trade-off between long and short local stories. The proposed formulation considers all segments equally, independently of their length.

[2] Remember here that the possible local stories correspond to all possible rendering strategies, i.e. included or excluded from the story, of the segment clips.

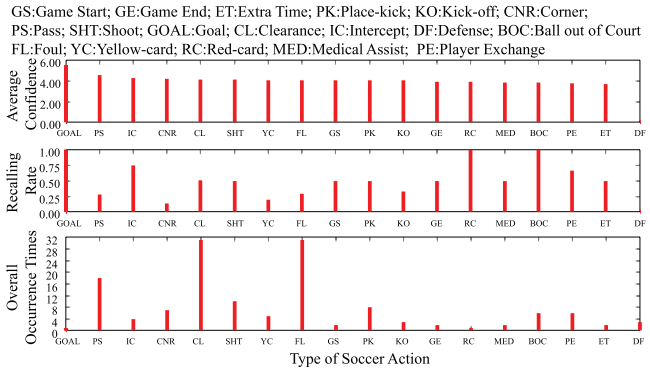
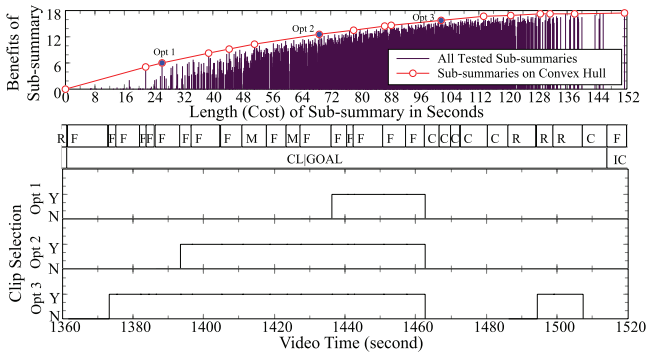


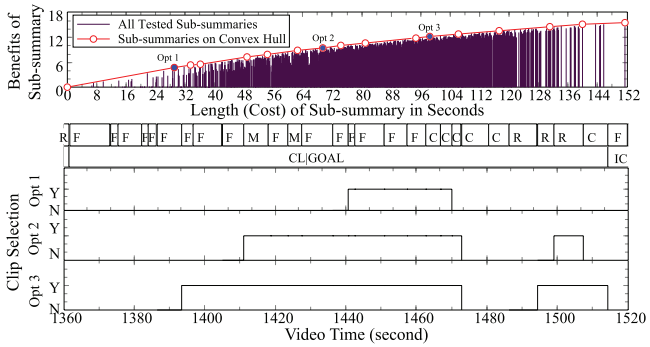
Fig. 4. Recall rates and level of significance obtained from audio analysis. The reference set of events has been extracted manually.

mainly by favoring continuous story and diffusing the audio significance according to view type structure (see Fig.3). In the second part of each figure, below the convex-hull, we present the organization of three convex-hull optimal local-stories. Those local stories have been selected at similar cost values in each figure, so that we can compare how α affects the inclusion of replay and close-view clips in the local story. In Fig.5(a), more importance is given to the game ($\alpha = 0.25$), which results in stories with less replays and close views, but longer game periods. In contrast, Fig.5(b) omits some parts of the game, so as to include close views and replays. This reflects the personalization capabilities of our resource allocation framework, for which adaptation of parameters effectively impacts the set of optimal local-stories that are considered for each segment.

Figure 6 compares the global behaviour of our summarization framework with the one of a rudimentary system that naively extracts key frames around audio highlighted moments. Specifically, the naive method applies Gaussian RBF Parzen window around each audio annotation, and sets the interest of each one-second frames slot to the maximum response resulting from the multiple annotations surrounding the slot. It then selects the slots in decreasing order of interest until we reach the length constraint. Clips for which less than half of the duration and less than five seconds have been selected are removed from the summary. In Figure 6, the comparison is performed using a 10 minutes summary length constraint. The first row presents the number of (annotated) segments that have been selected as a function of the parameters of the summarization system. We observe that our system primarily deals with annotated segments. This is not the case for the naive system, which ends up in selecting part of unannotated segments when the width of the parzen window increases. Moreover, it appears that our system mainly trades-off the duration and the number of rendered segments by playing on the continuity gain factor. In the second (third) row, the average occupancy (redundancy) evaluates the detail level of each selected segment (of replay and closeup views in the selected segment), which is computed by the percentage of time-length of the selected portion in the overall segment (of closeup and replay in the selected segment). On each graph, error bars represent the standard deviation, computed on the entire video. We observe that average occupancy is stable and subject to smaller standard deviation for the proposed method. It reflects the fact that our system assigns similar portion of contents but not similar length of contents to render the segment. This is in accordance with the footnote included at the end of Section 4.1. In contrast, the naive system allocates similar number of frames to each segment, whatever the content. Regarding the average redundancy, we mainly observe that our system can effectively control the level of redundancy by playing on the α parameter.



(a) $\alpha = 0.25$



(b) $\alpha = 0.50$

Fig. 5. Benefit-cost convex-hulls, and comparison of convex-hull optimal local-stories obtained with distinct α parameters. A single segment, and two α parameters (see Eq.3) are considered. On each figure, below the convex-hull, the first row defines the viewtype structure of the segment (Close, Far, Medium, or Replay views), the second row presents the manual annotations of the segment, and the three last rows define the selection status of the frames of the segment, i.e. whether they are part of the summary or not.

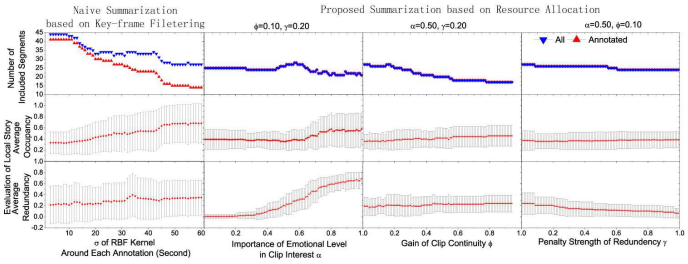
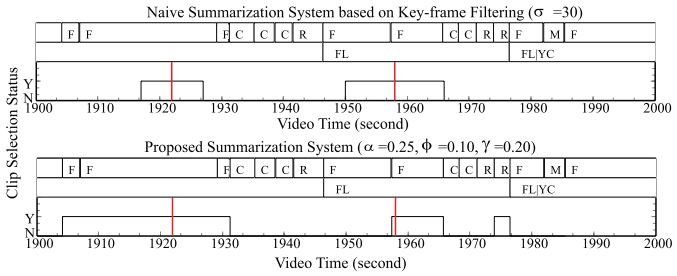
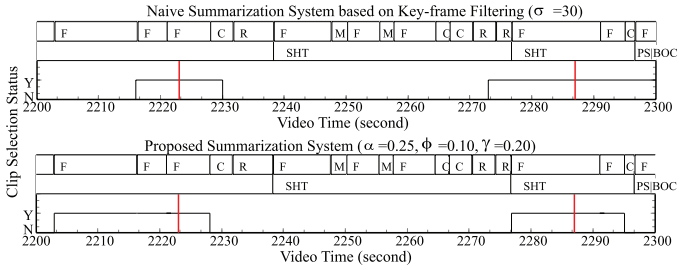


Fig. 6. Behaviour comparison between a rudimentary key-frame based summarization system, and the proposed framework.

To complete this section, we consider the two local summary video examples presented in Figure 7. In this figure, the first row of each graph defines how the segment is organized in close, far, and replay views. The second row defines the manual annotation of the segment, while red bars denote the automatically detected audio hot spots. Eventually, the third row identifies the frames that are selected to be included in the summary, respectively by the naive and proposed strategy. The supplementary material associated to this paper contains the videos corresponding to those segments. Grey scale images are used when the frame is not included in the summary, while color images are used to indicate inclusion in the summary. Those two examples demonstrate the benefit arising from the intelligent local story organization considered by our proposed framework. The first example corresponds to a case for which the audio hot-spot instant is somewhat displaced compared to the action of interest. As a consequence, the naive summarization system ends up in selecting frames that do not show the first foul action. In contrast, because it assigns clip interests according to view-type structure analysis, our system shows both fouls of the segment plus the replay of the second one. In the second example, the naive system renders the replay of the action that preceedes the action of interest, causing a disturbing story-telling artifact. In contrast, as a result of its underlying viewtype analysis and associated segmentation, our system restricts the rendering period to the segment of interest, and allocates the remaining time resources to another segment. Those examples are presented in details on the web page associated to this paper[13]. They clearly illustrate the benefit of our segment-based resource allocation framework.



(a)



(b)

Fig. 7. Two examples of the benefit arising from intelligent local story organization: (a) The segment includes two fouls. The naive summarization system ends up in selecting frames that do not show the first foul action. In contrast, our system shows both fouls and the replay of the second one. (b) The naive summary includes a replay of an action that is not rendered. Our system restricts rendering to the segment of interest, and allocates the remaining resources to another segment.

6. CONCLUSIONS

This paper has presented an effective and generic framework for personalized summarization of broadcasted soccer games. The main contributions consists in (1) an audio analysis tool to detect hot spots along the audio feed of the game, (2) a strategy to segment the game into semantically meaningful segments based on video shot detection and viewtype analysis, and (3) a resource allocation framework to select the video shots to include in the summary, as a function of individual user preferences. As a consequence, the proposed framework avoids setting the automatic recognition of events as a prerequisite task of summarization. It is however flexible enough to take advantage of any additional semantic information, such as the one provided by the latest achievements in automatic scene understanding, to further increase the personalization of user experiences. For content providers, our system offers the capability to deploy additional services to achieve more profits in a cost-efficient way. For customers, it offers personalization capabilities, e.g. for mobile or internet accesses.

7. REFERENCES

[1] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[2] D. Sadlier, N. O'Connor, N. Murphy, and S. Marlow, "A framework for event detection in field-sports broadcasts based on svm generated audio-visual feature model," in *IWSSIP'04*, Poznan, Poland, September 2004.

[3] B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," in *ICASSP*, Hong-Kong, April 2003, pp. 169–172.

[4] B.W. Chen, J.C. Wang, , and J.F. Wang, "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Trans. on Multimedia*, vol. 11, no. 2, pp. 295–312, Feb. 2009.

[5] Baillie M. and Jose J.M., "Audio-based event detection for sports video.," in *Proc. 2nd Int. Conf. Image video Retrieval, Lecture Notes in Computer Science, Urban, IL*, July 2003, vol. 2728, pp. 300–309.

[6] J. Li, T. Wang, W. Hu, M. Sun, and Zhang Y., "Soccer highlight detection using two-dependence bayesian network.," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo (ICME'06).*, July 2006, pp. 1625 – 1628.

[7] H. Duxans, X. Anguera, and D. Conejero, "Audio based soccer game summarization," in *IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting*, Bilbao, Spain, May 2009.

[8] J. Owens, "Tv sports production," *Focal Press*, 2007.

[9] Fan Chen and C. De Vleeschouwer, "A resource allocation framework for summarizing team sport videos," in *IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009.

[10] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. on Signal Processing*, vol. 36, no. 9, pp. 1445–1453, Sept. 1988.

[11] Antonio Ortega, "Optimal bit allocation under multiple rate constraints," in *Data Compression Conference*, Snowbird, UT, April 1996, pp. 349–358.

[12] D. Delannay, C. De Roover, and B. Macq, "Temporal alignment of video sequences for watermarking systems," in *Proc. SPIE*, 2003, vol. 5020, pp. 481–492.

[13] See additional results on the web, "http://www.jaist.ac.jp/ chenfan/apidis/www/icme-results.htm," 2010.