Formulating Team-Sport Video Summarization as a Resource Allocation Problem

Fan Chen and Christophe De Vleeschouwer

***Abstract*—We propose a flexible framework to summarize team-sport videos that have been originally produced for broadcast purposes. The framework is able to integrate both the knowledge about displayed content (e.g. level of interest, type of view, etc.), and the individual (narrative) preferences of the user. It builds on the partition of the original video sequence into independent segments, and creates local stories by considering multiple ways to render each segment. We discuss how to segment videos automatically based on production principles, and design parametric functions to evaluate the benefit of various local stories from a segment. Summarization by selection of local stories is then regarded as a resource allocation problem, and Lagrangian relaxation is performed to find the optimum. We investigate the efficiency of our framework by summarizing soccer, basketball and volleyball videos in our experiments.**

***Index Terms*—Sport Video Summarization, Content Repurposing, Resource Allocation**

I. Introduction

This paper considers sport video event summarization. The purpose is the generation of a concise video with well-organized and personalized story-telling. In contrast to many works devoted to the automatic detection of key actions, e.g. in football games [1][2][3] or to the compaction/reduction of video sequences for efficient browsing purposes[16]-[23], less 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 a list of highlighted actions, most earlier methods just implement pre-defined filtering or ranking procedures to extract some portions of the audiovisual stream that surround the actions of interest. 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, they just arbitrarily extract 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 storytelling artifacts. Our work attempts to overcome those limitations by introducing a generic resource allocation framework to adapt the selection of audiovisual 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 state-of-the-art, as highlighted in Section IV.

Our approach focuses on semantic story-telling rather than on video understanding. Given a set of sparse and instantaneous annotations about the scene at hand, a first contribution of our work consists in defining how to delimitate and organize local stories to render important game actions, in a semantically meaningful way, e.g. so that the beginning of an action is not shown without its end. The second major contribution of the paper lies in the definition of multiple local stories for each action, and in the selection of the local options that best match users preferences, both in terms of narrative style and semantic interest, under a global summary duration constraint.

This is achieved by analyzing the view structure of the broadcasted video at the light of general principles of sport video production, and by designing benefit/cost functions to evaluate how the different segments of the video contribute to the user satisfaction. We then solve summarization as a constrained resource allocation problem by defining a set of candidate local stories for each video segment, and propose a novel framework to select and organize those local stories into sport video summaries.

When all meta-data are effortlessly recordable from the production process, for content providers, the proposed framework offers the capability to deploy additional services to achieve more profits in a cost-efficient way. For customers, it offers personalization of access. The more accurate and abundant the meta-data, the more personalized the user experiences. When annotations are not directly provided by the production room, they can be generated automatically based on a proliferation of algorithms dedicated to key event detection and/or recognition, as surveyed in Section IV.

The paper is organized as follows. In Section II, we explain the basic idea of solving the summarization problem based on a resource allocation framework. We first list the set of required meta-data (such as shot-boundaries and view types) and their acquisition in Section II-A. In Section II-B, we propose the way to segment sport videos according the sequence of shots view types, by exploiting in a reverse engineering way the principles underlying sport video production. For each video segment, various candidate sub-summaries are considered, and resource allocation mechanisms are designed to drive

E-mail:fan.chen.ucl@gmail.com,christophe.devleeschouwer@uclouvain.be.
This work has been partly funded by FP7 Apidis European Project, DGTRE Walcomo project, and Belgium NSF.

the selection of the sub-summary that best meet end-users expectations in terms of content and story-telling, as explained in details in Section II-C. Experimental results are given in Section III to investigate the behavior of our framework, and evaluate the performance of our implementation. In Section IV, we briefly review the literature and further clarify the advantages of the proposed method. Finally, we conclude this paper and briefly explore our future work in Section V.

II. Summarization Through Resource Allocation

We first present an overview of the overall framework of our summarization method, which treats video summarization as an optimization problem to find the strategy of clip selection that achieves the highest benefit under the constrained summary length.

This paper takes as input the content that has been broadcasted by the production room. Figure 1 explains how personalized summaries are constructed based on the fragmentation of this content into non-overlapping and semantically meaningful segments. Especially, with all shot boundaries known, we divide a video into a sequence of clips of different view-types. According to production principles of sport videos[4], all non-replay clips in a sport video are temporally disjoint, and no dramatic camera switching occurs during a critical action. Consequently, we envision personalized summarization in the divide and conquer paradigm. By analyzing patterns of camera switching, we organize these clips into non-overlapping segments. Each segment corresponds to a short sub-story, which consists of consecutive clips that are semantically related. If we define a sub-summary, also named narrative option or local story, as one way to select clips within a segment, we regard the final summary as a collection of non-overlapping sub-summaries. Knowing view types of clips and events(-of-interest) in the video, all optimal combinations of clips within each segment are evaluated by their benefits and costs under specified user-preferences. We generate a universal set of candidate sub-summaries with various descriptive levels, and search for the best combination of sub-summaries which maximizes the overall benefit under user-preferred constraints.

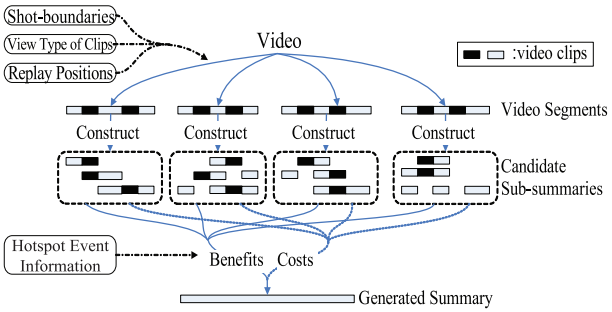

Fig. 1. Summarization in a divide-and-conquer paradigm.

We want to emphasize here that the proposed framework is generic in the sense that the benefits of each sub-summary can be defined in many ways, depending on the application context, the knowledge available or inferred about the scene at hand, and the narrative preferences of the user. Although we only investigate the efficiency using soccer, basketball and volleyball videos, the proposed method could also be extended to other team sports, with proper modification on the segmentation rule and benefit definitions.

The rest of this section explains how to acquire the meta-data, how to cut the video into segments, how to derive benefit and cost for the multiple narrative options of each segment, and how to solve the resource allocation problem by using Lagrangian Relaxation.

A. Meta-data Preparation

As depicted in Fig.1, two kinds of meta-data are required by our system: production actions(position of replays, shot-boundaries, view types) and event information. Production actions are exploited both to cut the video into semantically meaningful segments and to tune the benefits assigned to the various local stories of a segment. In contrast, event information acts at a global level, and directly impacts the (non)inclusion of segments in the summary. In practice, those information are either directly available from the production room, or can easily be inferred from the produced content, using state-of-the-art audio/video analysis tools. In our implementation, we detect replays from producer-specific logos[5], extract shot-boundaries with a detector that has been developed in [6] to better deal with smooth transitions, and recognize the view-type by using the method in Ref.[1].

Our framework supports event annotations in various forms, e.g., star rating, verbal descriptions, importance from the score/clocking information, and hotspots from audio commentary, etc., because they are equivalent after being further translated into game relevance and emotional relevance to identify the importance of actions.

For the soccer videos, we use detailed and manually annotated events, recorded in a format depicted in Fig.2, so as to demonstrate the efficiency of re-using the abundant annotation data from the media content provider. Note that our framework can however support automatically generated annotations, e.g., an extended version of our method using automatically detected hot-spots from audio signals has been explained in Ref.[7].

As for the basketball and volleyball videos, neither annotation files nor audio feeds are available. Considering that basketball and volleyball change their scores much more often than soccer, we infer the actions from the varying score and 24 second clocking(for basketball only), where scores are easily detectable via digital recognition from on-screen texts in pre-specified positions.

B. 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.3, we explain the segmentation rule that we envision for sport videos. It is worth noting that the proposed segmentation process only relies on

```
<Log Version="2.0.1">
  <Date>26-Oct-2008</Date>
  <TC>17:06:37:13 </TC>
  <DateUser>26-Oct-2008</DateUser>
  <TCUser>06:18:04:13 </TCUser>
  <TCTable>0</TCTable>
  <Description>remata fuera</Description>
  <TapeID>
  </TapeID>
  <InterestLevel>0</InterestLevel>
  <VideoFormat>1</VideoFormat>
  <Keywords>
    <Keyword Type="KEYWORD">Corner</Keyword>
    <Keyword Type="KEYWORD">Adriano Correia (6)</Keyword>
    <Keyword Type="KEYWORD">Remate</Keyword>
    <Keyword Type="KEYWORD">Luis Fabiano (10)</Keyword>
    <Keyword Type="KEYWORD">Replay</Keyword>
  </Keywords>
  <AssociatedClips>
    <AssociatedClip UmID="TV6QUKaW" />
  </AssociatedClips>
</Log>
```

Fig. 2. A sample fragment of annotation file of events that is available from the production-room.

information that are directly available from the production process. In other words, we do not assume any complex hand-made annotation process, or sophisticated automatic analysis of the video sequence to segment it. Rather, we exploit the fact that game state transitions motivate scene switching, and are thus reflected in the production actions, to segment the video based on the monitoring of production actions, instead of (complex) semantic scene analysis tools. In other words, generic sport production principles (like story telling continuity, see [4]) are here exploited to infer the status of the game based on the observation of production actions, as reflected by clip view type transitions. We now describe the process in more details.

A general diagram of state transition in one round of offense/defense is given in the left of Fig.3. Starting with a kick-off type action (tip-off or ball possession exchange in basketball, and serve in volleyball), the offensive side makes trials of score after several passing actions. This trial ends up in one of three possible results: scoring, being intercepted or being an opportunity. Before a new round, exceptional actions might happen, which include foul, medical assistance, and player exchange. We regard the state chain from the action-start to one of the results as a semantically complete segment.

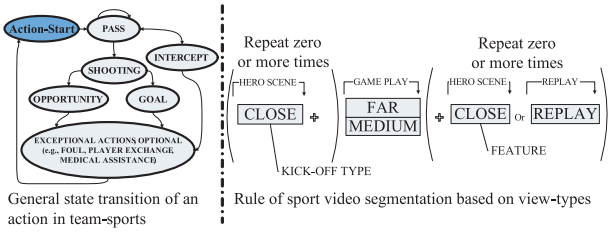

Fig. 3. Rule for segmentation of team-sports video.

In parallel, the graph on the right side of Fig.3 does present the sequential organization of video clips that results from the adoption of the team sport production principles [4] to deal with the game state transition observed in team sport actions. It depicts the typical view structure of a segment, and highlight how it is related to the semantic content of the scene. One segment usually starts with a close view for highlighting the player who kicks off. A sequence of far view and medium view will be the major part to tell the story of a segment. 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. Based on this structure, we divide the video into a series of segments. Although there are some complex cases, e.g., multiple successive trials of scoring or rapid revenge after successful interception, 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. When extra information is available, e.g., left/right court identification or 24-second clocking, these complex cases can also be further divided into finer segments. A similar analysis of the video in view-types was used in Ref.[3] to help detect exciting events(i.e., game parts with both close-up scenes and replays). In our case, we go one step further to infer the start and end times of an action based on the reverse analysis of production actions.

For completeness, we also note here that the extended version of our framework presented in [7], which summarizes soccer games by locating important actions based on the outcomes of an automatic analysis of the audio feed, does not rely on accurate prior annotation of actions, and has thus to be augmented by a kick-off detector. This is because, as highlighted in Fig.3, the correct segmentation of the video requires the identification of the close views that highlight a player just before a corner or a free kick. From the production perspective, the imminence of a kick-off is characterized by a still action, which in turns results in the absence of camera movement. Hence, the kick-off detector was simply and effectively implemented based on the measurement of the displacement of the field lines (as detected through chromatic analysis and Hough transform), compared to the camera viewport.

C. 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.,

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

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 (5s 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 utopian. 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 in Eq.(1).

To complete this section, we now explain how the significance associated to the 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 game relevance $\mathcal{I}_i^G$ and emotional level $\mathcal{I}_i^E$ for the i-th clip. The interest level $\mathcal{I}_i$ is then computed as

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

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

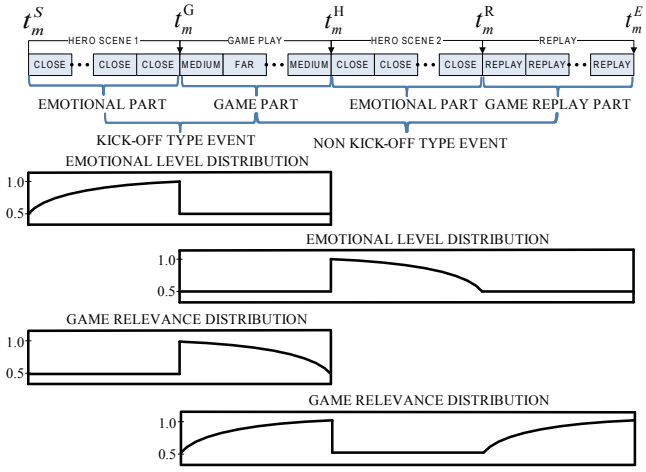

Fig. 4. Game relevance and emotional level are assigned as a function of clips view-types, where t_m^S, t_m^E are the starting and ending times of the m-th segment, and t_m^G, t_m^H, t_m^R are starting times of its game play part, hero scene part, and replay part, respectively.

Game relevance $\mathcal{I}_i^G$ and emotional level $\mathcal{I}_i^E$ of the i-th clip are computed by accumulating all related events for the m^{th} segment according to its view-type structure, i.e.,

$$\mathcal{I}_i^E = \mathcal{T}_m^E \sum_l \mathcal{D}_{li}^E \mathcal{G}_l^E, \qquad \mathcal{I}_i^G = \mathcal{T}_m^G \sum_l \mathcal{D}_{li}^G \mathcal{G}_l^G, \quad (4)$$

where $\mathcal{G}_l^E, \mathcal{G}_l^G$ represent the emotional level and game relevance assigned to the l-th highlighted event, respectively. $D_{li}^{E,G}$ denote the fraction of the emotional/game interest of the l^{th} event that is assigned to the i^{th} clip. We explain below how it is computed based on clips view-type knowledge. $\mathcal{T}_m^G$ 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]

As told above, $\mathcal{D}_{li}^G$ and $\mathcal{D}_{li}^E$ denote the percentage of game relevance and emotional level induced by the l-th event on the i-th clip, which satisfy $1 = \sum_i \mathcal{D}_{li}^E = \sum_i \mathcal{D}_{li}^G$. Motivated by considering production principles for team-sports, we design the distribution of game relevance $\mathcal{D}_{li}^G$ and emotional level $\mathcal{D}_{li}^E$ within a segment, as depicted in Fig.4.

- 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 events, we design their distributions based on the following facts: 1) Close views and replays are appended right after a critical action. A clip closer to

[1] Replacing $\mathcal{T}_m^G$ and $\mathcal{T}_m^E$ in Eq.(4) would enable extra controllability regarding the trade-off between long and short segments. The proposed formulation considers all segments equally, independently of their length.

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.

Especially, we use the hyperbolic tangent function to model the decaying process, because it is bounded and is integrable, which simplifies the computation of $\mathcal{D}_{li}^{E}$ and $\mathcal{D}_{li}^{G}$.

D. Global Summary Organization by Solving Resource Allocation Problem

We now explain how the global summary duration resource is allocated among the available local sub-summaries to maximize the aggregated benefit. Collecting at most one sub-summary for each segment, we form the final story. Here, in contrast to Section II-C, we refer explicitly to the segment index m, 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.,

$$\mathcal{B}(\{\mathbf{a}_{mk}\}) = \sum_{m} \mathcal{B}(\mathbf{a}_{mk}) \tag{5}$$

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.

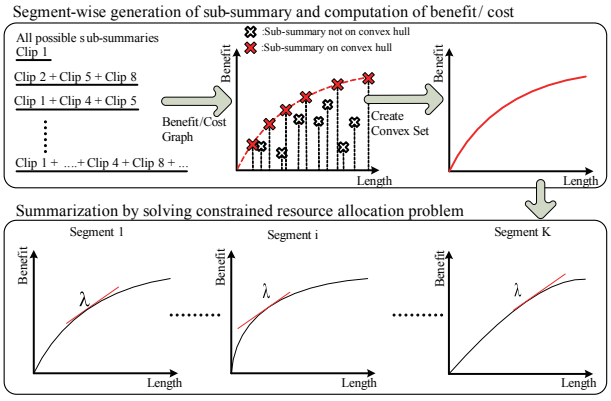

Fig. 5. Working flow in our summarization framework.

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 [9], [10]. 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 [9], [10]. 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 [11]. This results in a computationally efficient solution that can still consider a set of candidate sub-summaries with various descriptive levels for each segment. In Fig.5, a diagram summarizes the working process of summary organization by solving a resource allocation problem.

More specifically, we solve this resource allocation problem by using the Lagrangian relaxation [11], whose main theorem reads that if λ is a non-negative Lagrangian multiplier and $\{k^*\}$ is the optimal set that maximizes

$$\mathcal{L}(\{k\}) = \sum_{m} \mathcal{B}(\mathbf{a}_{mk}) - \lambda \sum_{m} |\mathbf{a}_{mk}| \tag{6}$$

over all possible $\{k\}$, then $\{\mathbf{a}_{mk^*}\}$ maximizes $\sum_m \mathcal{B}(\mathbf{a}_{mk})$ over all $\{\mathbf{a}_{mk}\}$ such that $\sum_m |\mathbf{a}_{mk}| \leqslant \sum_m |\mathbf{a}_{mk^*}|$. Hence, if $\{k^*\}$ solves the unconstrained problem in Eq.(6), then it also provides the optimal solution to the constrained problem in Eq.(5), with $u^{\mathrm{LEN}} = \sum_m |\mathbf{a}_{mk}^*|$. Since the contributions to the benefit and cost of all segments are independent and additive, we can write

$$\sum_{m} \mathcal{B}(\mathbf{a}_{mk}) - \lambda \sum_{m} |\mathbf{a}_{mk}| = \sum_{m} (\mathcal{B}(\mathbf{a}_{mk}) - \lambda |\mathbf{a}_{mk}|). \tag{7}$$

From the curves of $\mathcal{B}(\mathbf{a}_{mk})$ with respect to their corresponding summary length $|\mathbf{a}_{mk}|$, the collection of points maximizing $\mathcal{B}(\mathbf{a}_{mk}) - \lambda |\mathbf{a}_{mk}|$ with a same slope λ produces one unconstrained optimum. Different choices of λ lead to different summary lengths. If we construct a set of convex hulls from the curves of $\mathcal{B}(\mathbf{a}_{mk})$ with respect to $|\mathbf{a}_{mk}|$, we can use a greedy algorithm to search for the optimum under a given constraint u^{LEN}. The approach is depicted in Fig.5 and explained in details in [10]. In short, for each point in each convex hull, we first compute the forward (incremental) differences in both benefits and summary-lengths. We then sort the points of all convex-hulls in decreasing order of λ, i.e., of the increment of benefit per unit of length. Given a length constraint u^{LEN}, ordered points are accumulated until the summary length gets larger or equal to u^{LEN}. Selected points on each convex-hull define the sub-summaries for each segment.

In Table I, we list all operational parameters of the proposed framework. As a basic setting, the users are allowed to give their preferences on the length, the amount of replays and closeup views, and the continuity of the story. If the available annotations provide clear descriptions on the events or the dominant players (which are in fact available in the manual annotation from the production room), the users could further give their preferences on their favorite players and actions by tuning $\mathcal{G}_l^G$ and $\mathcal{G}_l^E$, which forms an extensive setting.

TABLE I
LIST OF ADJUSTABLE PARAMETERS FOR PERSONALIZED SUMMARIZATION

Personalization Ability of Our System					
	Basic Setting				Extensive Setting
Parameter	u^{LEN}	α	ϕ	γ	$\mathcal{G}_l^G, \mathcal{G}_l^E$
Dynamic Range	$[0, \infty)$	$[0, 1]$	$[0, \infty)$	$(-\infty, \infty)$	$[1, 5]$
Definition	Below Eq.(5)	Eq.(3)	Eq.(1)	Eq.(2)	Eq.(4)
Higher Value	Longer Summary	More Emotional Moments	More Continuous Story	Less Replays	Higher Interest
Lower Value	Shorter Summary	More Game Plays	Less Continuous Story	More Replays	Lower Interest

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

We have considered three different team sports videos to assess our summarization method: a whole soccer game along with a list of 200 annotated events from the production room in a format as shown in Fig.2, a basketball video and a volleyball video with events inferred from the score/clocking information. We only provide some representative results in the present paper. Their corresponding videos and more experimental results are available on the web page associated to this paper in Ref.[12].

We first investigate the behavior of our system, and then consider a set of subjective experiments to assess the solution with respect to actual feeds of end-users.

A. Behavior of the Summarization System

There are two major contributions of the proposed framework: view-type based segmentation which enables meaningful temporal partitioning for more efficient process without using complex semantic scene analysis, and high-level personalization of local-story organization through flexible benefit definition. We first verify the correctness of the proposed summarization framework in these two aspects.

In Fig.6, we show a part of segmentation results of all three videos. Dependent segments are merged into their previous segments. Inspecting the graph along with the videos in Ref.[12], we confirm that the view-type based segmentation is efficient in dividing videos into semantically intact actions for all tested team-sports. The view-type structure also reflects the pace of different team sports.

End users are allowed to personalize the summarization by specifying their interest on each type of event in terms of game relevance and emotional level[2], and setting their preferences on four parameters, i.e., summary length u^{LEN}, balance between game relevance and emotional level α, smoothness gain ϕ, and redundancy penalty γ. In Fig.7, we compare benefit-cost convex-hulls under four different parameter setups on the same segment of the soccer game. In each individual figure, we first plot at the top the set of (benefit, cost) pairs for all possible local stories of the segment[3] for a given configuration of parameter values. Convex-hull optimal sub-summaries are denoted by circles lying on the upper convex-hull of all (benefit, cost) points. Vertical lines represent other

[2]In practice, one can define a limited number of user profiles, each of them defining the levels of interests for a representative set of users.

[3]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.

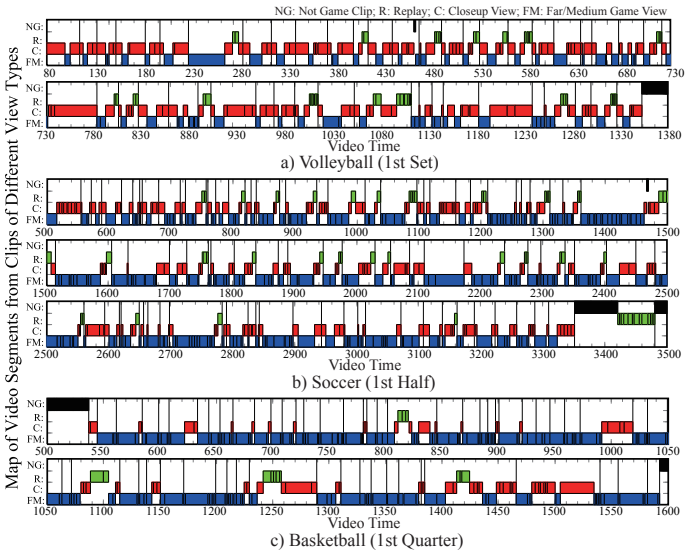

Fig. 6. Results of view-type based video segmentation for the soccer, basketball and volleyball videos in our experiments.

sub-optimal local stories. Below the convex-hull, we present the organization of three convex-hull optimal local-stories being selected at similar cost values (same duration) in each figure. 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, mainly by favoring continuous story and diffusing the event interests according to view type structure (in Fig.4). This reflects the personalization capabilities of our resource allocation framework, where adaptation of parameters effectively impacts the set of optimal local-stories.

B. Comparative Evaluation

As explained above, our framework aims at focusing on semantically relevant and personalized story telling. Due to its design principles, our proposed framework is also supposed to be efficient in stabilizing story organization, thereby improving robustness against temporal biases of (automatically detected) annotations. Those properties are explored through a comparative analysis with state of the art methods. Especially, we compared the behavior of our proposed method to the following three methods:

- a naive key-frame filter, which extracts the frames lying around pre-specified hot-spots. This naive method applies a Gaussian RBF Parzen window (σ being the standard deviation) around each hot-spot annotation, and sets the interest of each frame to the maximum response resulting from the multiple annotations surrounding the frame. It

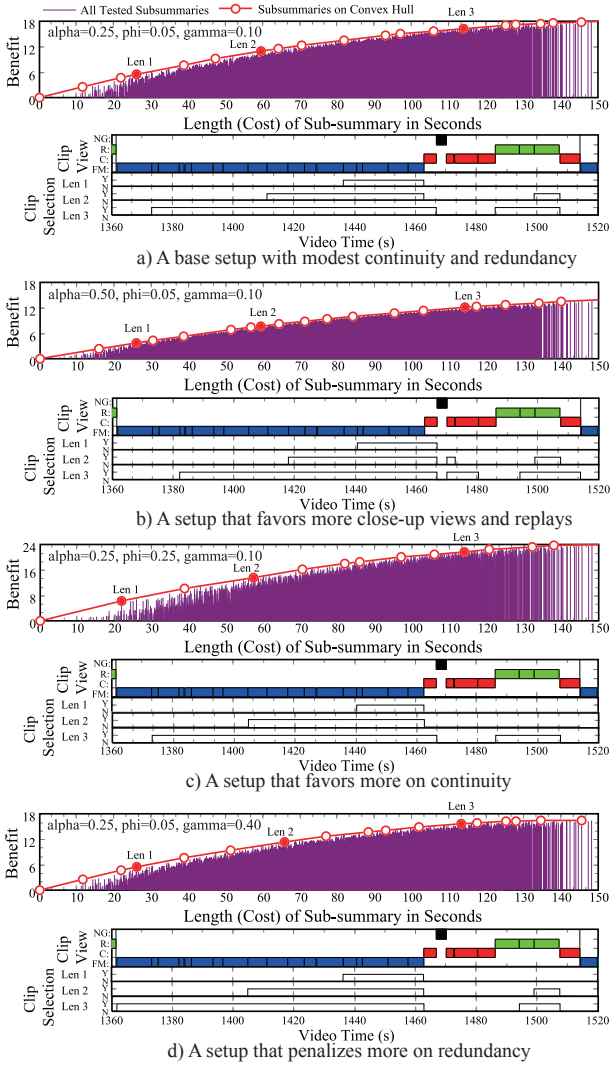

Fig. 7. Various organization of local stories under 4 different personalization options.

then selects the frames in decreasing order of interest until reaching the length constraint.

- LIU2007[14], which uses dynamic programming to locate both key-frames and their corresponding shot-boundaries simultaneously, by minimizing the reconstruction error between the source video and the extracted key-frames.
- LU2004[15], which selects pre-detected shots by maximizing the accumulated mutual distance between consecutive shots in the summary, subject to a length constraint. The mutual distance is evaluated from both histogram difference and temporal distance.

Instead of using manual annotations, we use automatically detected hotspots from audio commentaries, to investigate the robustness against biases/errors of event annotations. (Explanations on audio hot spot detection could be found in [7]).For computational efficiency, each key-frame is represented by a 1s temporal slot instead of a single moment in the above key-frame extraction methods, where computation defined on each key-frame is performed on the first frame of each 1s slot. From a portion of the source video, i.e., from 1300s to 2300s, each method is asked to organize a 150s summary.

Those resulting summaries are plotted in Fig.8. 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 vertical bars denote the automatically detected audio hot spots. Eventually, the following four rows identify the temporal occupancy of the summaries generated by the four tested methods. The web page associated to this paper [12] contains the videos corresponding to those segments. We made the following major observations:

1) Personalization Capability WRT Semantic User Preferences. Although the independence to annotations could be regarded as an advantage of LIU2007 and LU2004, it is rather obvious from their produced summaries that these two methods failed to reflect the relative importance of semantic events. Furthermore, they do not offer any flexibility to personalize the summary to satisfy semantic user preferences. The naive filter and the proposed method depend on annotated highlight events, and are thus able to tune the stories according to the relative importance of events.

2) Personalization Capability WRT Narrative User Preferences. It appears that both LIU2007 and LU2004 penalize the production of long-term continuous contents. LIU2007 is designed to include short representative key-frames, sparsely and evenly distributed in the whole video. LU2004 discourages selecting consecutive shots to maximize the mutual distance, and favors close-up/medium views and replays over far views because they contain more histogram variances than far-view grasslands. Furthermore, since the mutual distance is defined to be independent from the shot length, LU2004 also favors including more short shots over including less long shots, so as to maximize the accumulated mutual distance. However, in team sport videos, long-term far views are essential for the audience to understand the complexity of the teamwork. Continuity is also important in telling a fluent story. Frequent switching between short clips leads to very annoying visual artifacts in their corresponding video data. Finally, neither LIU2007 nor LU2004 has the flexibility to adjust the story-telling pattern.

In contrast, our proposed method and the naive key-frame filter favor continuous contents under certain parameter settings. The naive filter always expands the summary around the given annotation, which makes it sensitive to biased annotations. The proposed framework further considers intelligent organization of replays and different view-types to satisfy various narrative user preferences.

3) Robustness Against Annotation Error. The proposed method has improved robustness against temporal biases of (automatically detected) annotations. Two examples demonstrate the improved robustness arising from the intelligent local story organization considered by our proposed framework. The first example (1900s-1940s) 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 key-frame filtering 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 (2260s-2300s), the naive system

renders the replay of the action that precedes the action of interest, causing a disturbing story-telling artifact. In contrast, as a result of its underlying view-type analysis and associated segmentation, our system restricts the rendering period to the segment of interest, and allocates the remaining time resources to another segment.

All these results clearly illustrate the benefit of our segment-based resource allocation framework with respect to story-telling and personalization capabilities.

In Table II, we give the computation time of the four investigated methods. LIU2007 is very time consuming, although we have taken several optimization techniques, e.g., using look-up tables to host precomputed features. If we perform a finer selection on each frame instead of the above 1s frame slot and deal with the whole video, it could cost hours to produce the summary (the complexity is $O(kn^3)$ for extracting k key-frames from n frames in Ref.[14]), which makes it less practical in a real applicative scenario. In the other three methods, it costs at most a few seconds to produce the video, which is thus able to give a real-time response to users.

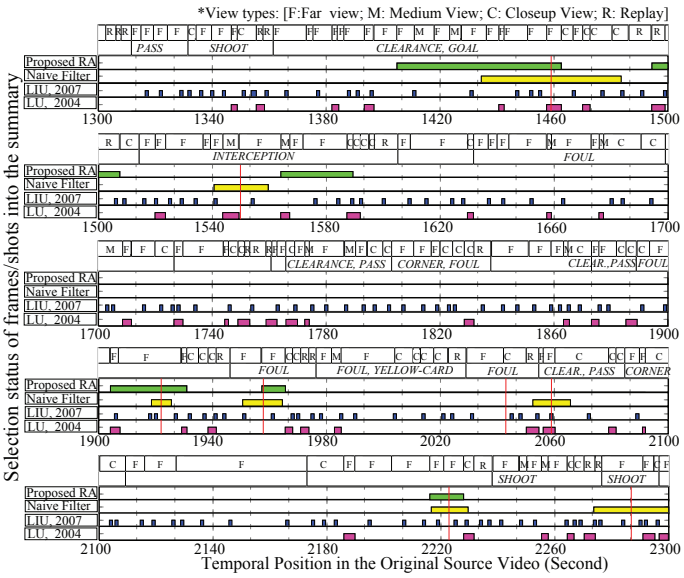

Fig. 8. We plot the summaries resulting from four different methods. The first row of each graph presents the view-structure of segments. The second row defines the manual annotation of the segment, while red vertical bars denote the automatically detected audio hot spots. Eventually, the following four rows identify the temporal occupancy of the summaries generated by the four tested methods. (We set $\sigma = 30$ for the naive kf filter, and used $\alpha = 0.25, \phi = 0.10, \gamma = 0.20$ for the proposed method.)

TABLE II
TIME USED IN SUMMARIZING THE 1000S LONG SOCCER VIDEO

Proposed RA	Naive KF Filter	Liu, 2007	Lu, 2004
0.068s	0.003s	435.3s	2.5s

C. Results of Subjective Evaluation

Different purposes of video summarization lead to different evaluation methods. In contrast to those methods which regard video summarization as maximizing a pre-defined similarity between the summary and the original video[14][15], our method cares more about satisfaction of user preferences and story-telling. Hence, we rely on subjective tests to evaluate the performance of our summarization system.

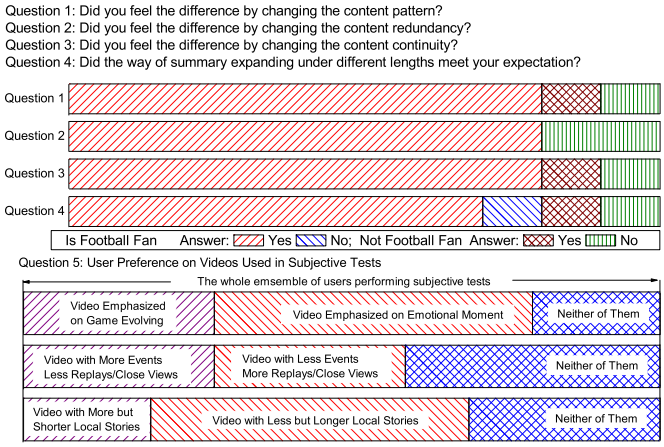

Fig. 9. Answers from 10 viewers who performed subjective evaluation of our summarization system after viewing several pairs of summarized videos.

We prepared a web page for onlined subjective evaluation. The viewers are asked to view several pairs of summarized videos generated under different parameters, and then to answer some questions about their appreciation. Four groups of data have been generated: one group of two 6-minute videos focusing on either explaining game evolving (Less Replays/Close Views, More Events) or increasing emotional involvement (More Replays/Close Views, Less Events), one group of two 6-minute videos with different content redundancy, one group of two 6-minute videos with different content continuity, and one group of four videos with different summary lengths ($1, 3, 6, 12$ minutes) but identical narrative parameters. After viewing each group, the viewer is asked whether the difference is visually perceptible or not. Furthermore, for the top three groups, we asked the viewer to write down his preference on one of the two videos and his free comments (e.g., any artifacts he saw). We summarize the answers from 10 viewers, and plot them in Fig.9. From question 1 to 4, we confirm that most of the viewers (especially soccer fans) could visually discriminate these videos produced with different parameter values, and were satisfied by the way our system did expand the story according to various summary lengths. From the preferences they expressed within each group of summarized videos, we draw two major conclusions:

Personalization appears to be relevant in the sense that (1) viewers have a preference, and (2) there is not a clear consensus about the preferred option among football fans. Note also that among the people who prefer neither of the videos, half of them do not like soccer.

From the results, it appears that more users prefer a story detailing a few events with related emotional moments, such as replays and close-up views. Less users prefer a brief global review of the game. Therefore, it is a priori more important to tell the story of each action extensively than to list out all highlighted events. However, this is not true for all users, since about 30% of football fans did prefer to increase the amount of local stories.

As a conclusion, those subjective results confirm that different users may have different viewing purposes, which reinforces our motivation to abstract summarization as a trade-off

TABLE III
LIST OF DETECTED ARTIFACTS IN ALL THREE SUMMARIES AND THEIR POSSIBLE REASONS

Artifact Index	*Related Clips*	*Category**	*Observation*	*Likely Reason*
Summary-1: 7 significant artifacts from 17 clip gaps (Occurrence Rate 41.2%)				
1	92, 98	3	92: Ball out of border; 98: Corner kick-off	Viewers need a break to follow the action
2	100, 148	3	100: Opportunity; 148: Place-kick	Missing action (Cause of Place-kick)
3	240, 247	1	240: Terminated passing;	Biased shot-boundary
4	444, 481	3	444: Goal keeper got the ball; 481: Place-kick	Missing action (Cause of Place-kick)
5	684, 728	3	684: Foul; 728: A long shooting	Missing action (Consequence of Foul)
6	729, 879	2	729: Goal keeper got the ball; 879: Place-kick	Incomplete action (879 without beginning)
7	1022	1	1022: Shooting without ending	Biased shot-boundary
Summary-2: 5 significant artifacts from 12 clip gaps (Occurrence Rate 41.7%)				
1	100, 148	3	100: Opportunity; 148: Place-kick	Missing action (Cause of Place-kick)
2	150, 175	3	150: Close-up view of opportunity; 175: Place-kick	Missing action (Cause of Place-kick)
3	240, 245	1	240: Terminated passing;	Biased shot-boundary
4	684, 728	3	684: Foul; 728: A long shooting	Missing action (Consequence of Foul)
5	1022	1	1022: Shooting without ending	Biased shot-boundary
Summary-3: 16 significant artifacts from 66 clip gaps (Occurrence Rate 24.2%)				
1	113, 124	3	113: Foul action; 124: Passing;	Missing action (Consequence of Foul)
2	128, 136	3	128: Replay of foul; 136: Another foul;	Missing action (Consequence of Foul)
3	161, 163	2	161: Replay; 163: Foul;	Incomplete action (163 without beginning)
4	169, 175	3	169: Replay of clearance; 175: Place-kick	Missing action (Cause of Place-kick)
5	240, 247	1	240: Terminated passing;	Biased shot-boundary
6	304, 308	3	304: Ball out of border; 308: Corner kick-off	Viewers need a break to follow the action
7	333, 337	2	333: Foul; 337: Replay of the foul	Viewers failed to catch the replay
8	337, 340	1	340: A short closeup of the goalkeeper included	Biased shot-boundary
9	453, 481	3	453: Clearance; 481: Corner-kick;	Missing action (Cause of Corner-kick)
10	621, 670	3	621: Goal keeper got the ball; 670: Rapid passing;	Viewers need a break to follow the action
11	771, 799	3	771: Foul; 799: Another foul	Missing action (Consequence of Foul)
12	811, 834	3	811: Short replay of foul; 834: Passing;	Viewers need a break to follow the action
13	856, 867	2	856: Clearance; 867: Foul;	Incomplete action (867 without beginning)
14	873, 879	2	873: Replay of foul; 879: Place-kick;	Incomplete action (879 without beginning)
15	1022, 1026	1	1022: Shooting without ending	Biased shot-boundary
16	1030, 1041	3	1030: Goal keeper got the ball; 1041: Other action;	Viewers need a break to follow the action

*Category 1: Story-telling artifacts due to erroneous or missing meta-data; 2: Story-telling artifacts due to incomplete local stories; 3: Story-telling artifacts due to discontinuity between two consecutive local stories.

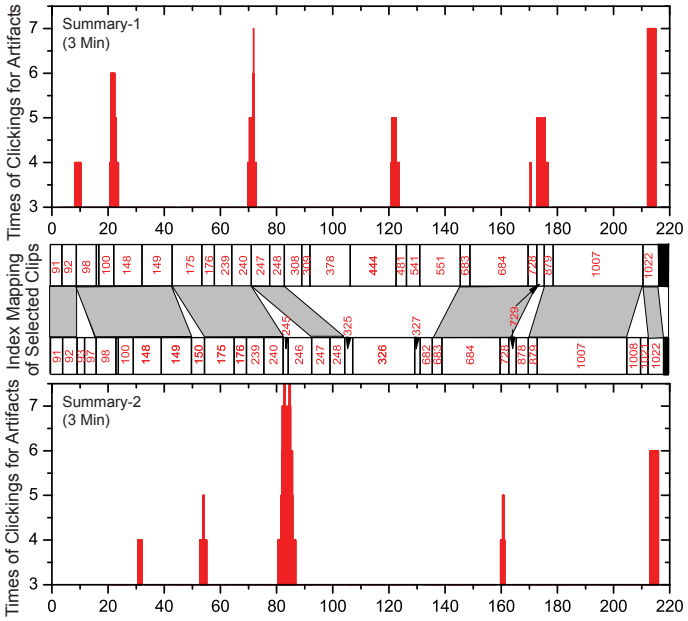

Fig. 10. Story-telling artifacts detected by 17 viewers on two 3 minute summaries. Each viewer is asked to see the video and click the mouse when he sees story artifacts. The density of clicking is estimated by using the Parzen-window function. A higher density means a higher probability of artifact occurrence.

between several key concepts (i.e., continuity and redundancy, game relevance and emotional level), and to focus on the organization of local stories as well as on that of the global story. Simple mechanisms of video summarization in the literature, such as linear-filtering of highlighted events or minimization of the reconstruction error between the summary and the original video, lack the flexibility to control organization of local stories according to user preferences.

This is confirmed by the second and independent subjective evaluation, which has been run to assess the relevance of our story organization. Through a tool we have developed, each viewer is asked to see the video and click the mouse when he sees any kind of story-telling artifacts. The timestamp of clicking is automatically recorded by the tool. We do not ask viewers to input detailed comments after each clicking, because interruption during video playing might distract viewers from focusing on the story evolving in the summary. As a consequence, we have to find out the reason behind each clicking by analyzing the aggregation of clickings, a posteriori. We estimate the density of clickings at each video time by using the Parzen-window function, where a rectangular window of width 3 seconds is applied to the left side of each clicking to compensate the delay between the occurrence of story artifacts and the corresponding clicking. We collected results from 17 viewers (15 males and 2 females, age 20 to 40).

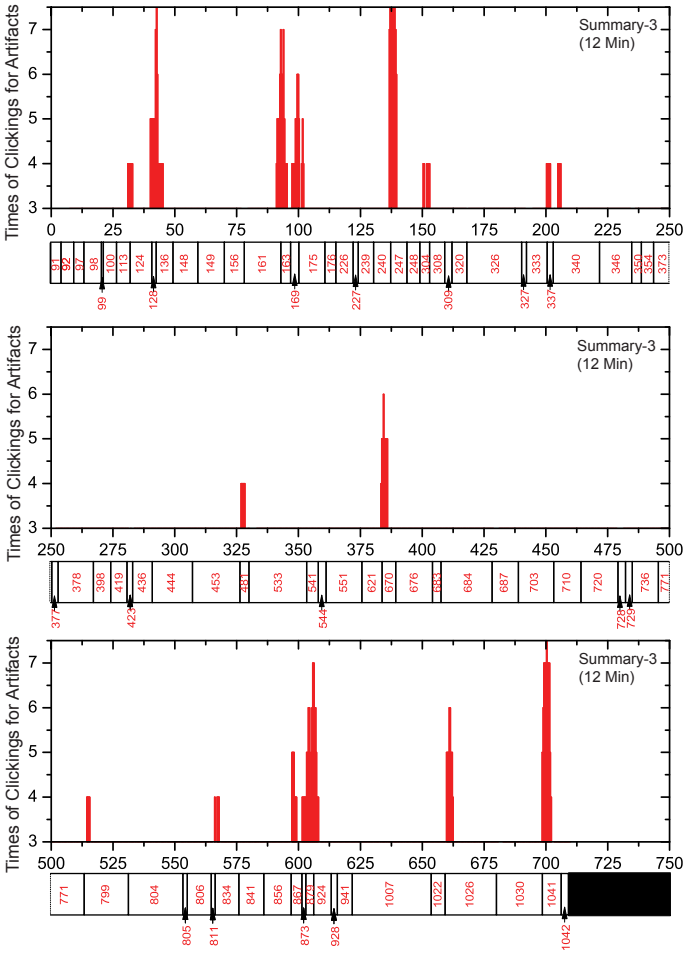

Fig. 11. Story-telling artifacts detected by 17 viewers on one 12 minute summary. Each viewer is asked to see the video and click the mouse when he sees story artifacts. The density of clicking is estimated by using the Parzen-window function. A higher density means a higher probability of artifact occurrence.

Three summaries have been considered in the test, with two of them being 3 minutes long and one 12 minutes. Clicking times of artifacts in the two 3-minute summaries are compared in Fig.10, and results on the 12-minute video are given in Fig.11. In Fig.10, we also give the corresponding index of selected clips in the summary and depict the mapping of clips between these two summaries. It is obvious from Fig.10 that summary 2 has less but longer local stories than summary 1.

Before giving a detailed analysis of each individual artifact, we need to mention that users were globally satisfied with our results. There are only a relatively small number of clicks, although there were numerous borders, for which users had many opportunities to click during the subjective test. Furthermore, there are less than two significant artifacts per minute. Here, we define a significant artifact as an artifact marked by more than three viewers. Also, there are few artifacts that have been pointed out by all viewers. Therefore, most artifacts are not severe, since they are not considered as being annoying by some of the viewers. The reviewers are invited to access summary samples at Ref.[12] to make their own opinions. A simplified onlined version of our summarization system is also accessible through our homepage [13].

Through careful analysis of significant artifacts pointed by the viewers, we list all significant artifacts and their possible reasons in Table III. We identify three main categories of artifacts as follows:

1) Story-telling artifacts due to erroneous or missing meta-data, such as biased shot-boundaries and incorrectly detected view types. These errors may cause the corresponding local story incomplete due to the missing of relevant frames or the inclusion of irrelevant frames, especially when they appear in the beginning or ending of a video segment. The only reliable way to remove artifacts in this category is to improve the accuracy of both shot-boundary and view-type detection.

2) Story-telling artifacts due to incomplete local stories. A common case is that the local story lacks of beginning or ending clip. Especially when an action (e.g., place-kick or free-kick) takes place at quite rapid pace, we need to include extra contents before the action, i.e. a close-up view of the dominant player who performs the place-kick, to guide viewers to reorient themselves in the new action.

3) Story-telling artifacts due to discontinuity between two consecutive local stories. Even though the local story is complete for presenting an action, it will be regarded as incomplete when the related actions explaining its reason or consequence are missing. Besides the artifacts explored in Table III, similar artifacts would also occur on events such as red/yellow card, medical assistance, and so on. Our framework can be naturally extended to remove this kind of artifacts, by including several video segments in the current system into a higher level of video segment, named ”super-segment”. Story organization in each ”super-segment” considers the dependency of closely related events in those segments. Some complete stories, e.g., the corner kick-off in clip 98 after ball getting out of the border, were marked as artifacts due to the difficulty in understanding the relationship at a glance. (When the viewer has built an overall idea about the game after viewing summary 1, no artifact has been reported in clip 98 in summary 2 and 3.) This observation reminds us the importance of breaks, such as close-up views and replays, in the original video, and drives us to consider their roles in the summary.

From Table III, artifacts in category 2 only occupy a small proportion of all artifacts, which confirms the efficiency of our approach in local story organization. Most of the artifacts come from category 3. Although ”super-segment” could be included to solve this problem, the system would then become severely dependent on the accuracy and completeness of the annotations. Hence, a balance that best fits the application context must be found.

IV. Related Works

Our approach focuses on presentation of video summaries, rather than extraction of events/objects-of-interest. Various methods were proposed according to their different summarization purposes [16][17]. Many works interpreted video summarization as extracting a short video sequence of a desired length, in a way that minimizes the loss resulting from the skipped frames and/or segments. Those methods differ in their various definition of the similarity between the summary and the original video, and in their diversified techniques to maximize this similarity. They include methods for clustering

similar frames/shots into so called key frames [18][19], and methods for constrained optimization of objective functions [20][21]. Refs. [1][22] measured precision and recall rates of three events in soccer game, which also belong to this category. Since they attempt to preserve as much as possible of the initial content, all those methods are well suited to support efficient browsing applications. For personalized browsing, summarization was also implemented as a "query" function to extract objects/events preferred by the user, via textual descriptors and with/without interaction [23] [24].

However, the motivation of end-users in viewing summaries is not limited to fast browsing of all clips in the whole video content. For instance, summarization in Ref.[25] aims at organizing a music soccer sport video, where synchronization between the video contents and the music contents is their major topic. A summary could also be organized to provide information to special users, e.g., helping the coach to analyze the behavior of players from their trajectories[26]. In the present paper, we intend to provide a concise video with well-organized story-telling, from which the audience could enjoy the highlight moments of all events that best satisfy their interest. In Ref.[27], continuity of clips in generated summary was considered for better story-telling. Ref.[28] organizes stories by considering a graph model for managing semantic relations among concept entities. Compared to the scenes for general videos, stories in sport videos have much simpler structures and a limited set of possible events. However, by investigating the literature, we found that personalization of story organization in previous methods for sport video summarization was mainly focused on assignment of event significance [29] and extraction of specified view types [3], according to user preferences. For a sport video with many overlapped events, these methods based on linear filtering of interests might produce a summary consisting of partial stories of those events without the most highlighted moment, because the overlapped part could have higher interest than the highlighted moment due to the accumulation of multiple events. In contrast, we prefer to present to the user a summary that explains any individual event completely. Furthermore, we target at personalized patterns of story-telling (including continuity, redundancy, and prohibited story organization, etc.) as well as personalized retrieval of events-of-interest. All these require further control of both local and global story-telling.

Utilization of event annotations from the production room might raise concerns on the practicality of our system. In fact, there are many methods in the literature that avoid the difficulty of automatic detection of sport events by pure image analysis, by directly using meta-data of events [18][24], or indirectly using meta-data of events, such as methods based on analyzing web-casted text data of sport events in Ref.[30], and methods that extract text data on events, scores and other semantic information directly from the video[29] [31]. As we explained in Section II-A, what really matters to the proposed framework is the significance of the event in the sense of game relevance and emotional level. This semantic gap between the video and the significance of its corresponding events can be filled by finding different abstracted representation of the events, such as using motion intensity to reflect the activity [32], using general attentional model to evaluate the interest of the video content [33], establishing the link between low-level features and user excitement by observing the temporal behavior of selected audiovisual features [34][7]. Recent progresses in the field of social media, such as the success of YouTube or wikipedia, also provide us an alternative way to collect explicit meta-data directly from end-users.

Whilst the proposed resource allocation framework can also support the summarization of raw multi-view content [35], the present paper focuses on (audio-) video contents released by the production room, and completed by annotations of events-of-interest that are collected during the production process itself [8] or result from automatic analysis tools (e.g. the audio feed analysis tool considered in [7]).

Our proposed method has four major advantages: 1.) Highly flexible personalization. Typically, several production factors, e.g., continuity (summary with less but complete local stories or more but incomplete local stories) and redundancy (summary with more or less replays), collaborate with event significance to produce the final result. 2.) Improved story-telling complying with production principles of sport videos. One original contribution of our work is to infer the level of interest of clips by analyzing the role of each view type in sport production to present a sport event. 3.) Graceful handling of knowledge about events of interest. By analyzing the role of each shot in presenting a segment, our method can deal with temporally biased event annotations. 4.) Dynamic story. It searches for optimal combination of clips by considering the purpose of each type of views in story-telling, and provides dynamic stories with respect to various user preferences regarding narration and context. Here, we propose a framework that naturally and elegantly interprets non-linear and non-additive benefits arriving from the concatenation of clips describing (a part of) the same action. Hence, it goes much beyond approaches that simply filter out the segments or clips with too small interest level.

V. Conclusions

We proposed a framework for producing personalized summarization of sport videos. The video is divided into many short clips. Within each local time segment, we deal with short clips to carefully organize the story-telling in resource allocation framework. Globally, we use Lagrangian relaxation to find the optimal selection of clips to form the final summary. We thus consider the problem in the divide-and-conquer paradigm, which makes the current summarization system real-time with pre-computed meta-data. Our method is flexible in the sense that it supports different definition of benefit to customize the summarization process.

Subjective tests validate the approach and explored the direction of further improvements. We will further consider the dependency between closely related events to achieve better presentation of actions. We also need more subjective evaluations to define parameter profiles, thereby simplifying the task of the user. Inclusion of audio information is a central task in our near future, which brings not only benefits such as better user experiences and more information for scene understanding, but also challenges such as synchronization

between completeness of both video and audio summaries. Extension to other contexts, e.g., social media, cases for which none information is available about the game (Basketball scenario in APIDIS) will also be discussed.

REFERENCES

[1] Ekin A., Tekalp A.M., and Mehrotra R., "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing,* vol.**12** pp.796-807, 2003.

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

[3] Li B., Pan H., and Sezan I. "A general framework for sports video summarization with its application to soccer," *ICASSP'03,* vol.**3** pp.169-172, 2003.

[4] Owens J., "Television sports production," *Focal Press*, 2007.

[5] Pan H., van Beek P., and Sezan M.I., "Detection of slow-motion replay segments in sports video forhighlights generation," *ICASSP'01,* vol.3, pp.1649-1652, 2001.

[6] Fernandez I.A., Chen F., Lavigne F., Desurmont X., and De Vleeschouwer C., "Browsing Sport Content through an Interactive H.264 Streaming Session,", *MMEDIA 2010,* pp.155-161, 2010.

[7] Chen F., De Vleeschouwer C., Barrobes H.D., Escalada J.G., and Conejero D., "Automatic and personalized summarization of audio-visual soccer feeds," *ICME 2010,* pp.837-842, 2010.

[8] Chen F., De Vleeschouwer C., "A resource allocation framework for summarizing team sport videos," *ICIP 2009,* pp.4349-4352, 2009.

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

[10] Ortega A., "Optimal bit allocation under multiple rate constraints," *Data Compression Conference,* pp.349-358, 1996.

[11] Everett H., "Generalized lagrange multiplier method for solving problems of optimum Allocation of Resources," *Operations Research,* vol.**11** pp.399-417, 1963.

[12] Supplemental Materials:
http://www.jaist.ac.jp/~chen-fan/apidis/www/results-tcsvt-new.html

[13] Homepage of APIDIS project: http://www.apidis.org/
Online Summarization Demo:
http://www.apidis.org/asp/gensummary.asp

[14] Liu T. and Kender J.R., "Computational approaches to temporal sampling of video sequences," *ACM Trans. on Multimedia Computing, Communications, and Applications,* vol.2, 2007.

[15] Lu S., King I., and Lyu M.R., "Video summarization by video structure analysis and graph optimization," *ICME 2004,* pp.1959-1962, 2004.

[16] Money, A.G., and Agius, H, "Video summarization: a conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation,* vol.19, pp.121-143, 2008.

[17] Truong, B.T. and Venkatesh, S. "Video abstraction: A systematic review and classification," *ACM Transactions on Multmedia Computing, Communication and Application,* vol.3, 2007.

[18] Tseng B.L., and Smith J.R., "Hierarchical video summarization based on context clustering," *J.R. Smith, S. Panchanathan, T. Zhang (Eds.), Internet Multimedia Management Systems IV: Proceedings of SPIE,* vol. 5242, pp.14-25, 2003.

[19] Ferman A.M., and Tekalp A.M., "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Trans. on Multimedia,* vol.5, pp.244-256, 2003.

[20] Li Z., Schuster G.M., and Katsaggelos A.K., "MINMAX optimal video summarization," *IEEE Trans. on Circuits and Systems for Video Technology,* vol.15, pp.1245-1256, 2005.

[21] Pahalawatta P.V., Zhu L., Zhai F., and Katsaggelos A.K., "Rate-distortion optimization for internet video summarization and transmission," *MMSP 2005,* pp.1-4, 2005.

[22] Cai R., Lu L., Zhai H., and Cai L., "Highlight sound effects detection in audio stream," *ICME 2003,* pp.37-40, 2003.

[23] De Silva G., Yamasaki T., and Aizawa K., "Evaluation of video summarizationf or a large number of cameras in ubiquitous home," *ACM MM,* pp.820-828, 2005.

[24] Takahashi Y., Nitta N., and Babaguchi N., "Video summarization for large sports video archives," *ICME 2005,* pp.1170-1173, 2005.

[25] Wang J., Xu C., Chng E., Duan L., Wan K., and Tian Q., "Automatic Generation of Personalized Music Sports Video," *ACM Multimedia,* pp.735-744, 2005.

[26] Zhu G., Huang Q., Xu C., Rui Y., Jiang S., Gao W., and Yao H., "Trajectory Based Event Tactics Analysis in Broadcast Sports Video," *ACM Multimedia,* pp.58-67, 2007.

[27] Albanese M., Fayzullin M., Picariello A., and Subrahmanian V.S., "The priority curve algorithm for video summarization," *Information Systems,* vol.31, pp.679-695, 2006.

[28] Chen B.W., Wang J.C., and Wang J.F., "A novel video summarization based on mining the story-structure and semantic relations among concept entities," *IEEE Trans. on Multimedia,* vol.11, pp.295-312, 2009.

[29] Babaguchi N., Kawai Y., Ogura T., Kitahashi T., "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Trans. on Multimedia,* vol.6, pp.575-586, 2004.

[30] Refaey M.A., Abd-Almageed W., and Davis L.S., "A logic framework for sports video summarization using text-based semantic annotation," *SMAP 2008,* pp.69-75, 2008.

[31] Tjondronegoro D., Chen Y., and Pham B., "Highlights for more complete sports video summarization," *IEEE Trans. on Multimedia,* vol.11, pp.22-37, 2004.

[32] Bezerra F.N., and Lima E., "Low cost soccer video summaries based onvisual rhythm," *ICML 2006,* pp.71-77, 2006.

[33] Ma Y., Hua X., Lu L., and Zhang H., "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia* vol.7, pp.907-919, 2005.

[34] Hanjalic A., "Adaptive extraction of highlights from a sport video based on excitement modeling," *IEEE Trans. on Multimedia* vol.7, pp.1114-1122, 2005.

[35] Chen F., De Vleeschouwer C., "Automatic Production of Personalized Basketball Video Summaries from Multi-sensored Data," *ICIP 2010,* vol.1, pp.565-568, 2010.

Fan Chen received the BS degree in computer science from Nanjing University in 2001. He received the MS degree in information science from Tohoku University in 2005 and Ph.D. from Japan Advanced Institute of Science and Technology in 2008. He was a post-doctoral researcher in TELE, UCL, and worked for the FP7 APIDIS European project (2008-2010). He is currently an assistant professor in Japan Advanced Institute of Science and Technology. His research interests are focused on statistical inference and optimization techniques related to computer vision, pattern recognition and multimedia analysis.

Christophe De Vleeschouwer is a permanent Research Associate of the Belgian NSF and anAssistant Professor at UCL. He was a senior research engineer with the IMEC Multimedia Information Compression Systems group (1999- 2000), and contributed to projects with ERICSSON. He was also a post-doctoral Research Fellow at UC Berkeley (2001-2002) and EPFL (2004). His main interests concern video and image processing for communication and networking applications, including content management and security issues. He is also enthusiastic about non-linear signal expansion techniques, and their use for signal analysis and signal interpretation. He is the co-author of more than 20 journal papers or book chapters, and holds two patents. He serves as an Associate Editor for IEEE Transactions on Multimedia, has been a reviewer for most IEEE Transactions journals related to media and image processing, and has been a member of the (technical) program committee for several conferences, including ICIP, EUSIPCO, ICME, ICASSP, PacketVideo, ECCV, GLOBECOM, and ICC. He is the leading guest editor for the special issue on Multicamera information processing: acquisition, collaboration, interpretation and production, for the EURASIP Journal on Image and Video Processing. He contributed to MPEG bodies, and several European projects. He now coordinates the FP7-216023 APIDIS European project (www.apidis.org), and several Walloon region projects, respectively dedicated to video analysis for autonomous content production, and to personalized and interactive mobile video streaming.