An autonomous framework to produce and distribute personalized team-sport video summaries: a basket-ball case study

Fan Chen, Damien Delannay, and Christophe De Vleeschouwer

***Abstract*—Democratic and personalized production of multimedia content is a challenge that content providers will have to face in the near future. In this paper, we address this challenge by building on computer vision tools to automate the collection and distribution of audiovisual content. Especially, we proposed a complete production process of personalized video summaries in a typical application scenario, where the sensor network for media acquisition is composed of multiple cameras, which, for example, cover a basket-ball field. Distributed analysis and interpretation of the scene are exploited to decide what to show or not to show about the event, so as to produce a video composed of a valuable subset of the streams provided by each individual camera. Interestingly, the selection of the streams subsets to forward to each user depends on his/her individual preferences, making the process adaptive and personalized. The process involves numerous integrated technologies and methodologies, including but not limited to automatic scene analysis, camera viewpoint selection, adaptive streaming, and generation of summaries through automatic organization of stories. The proposed technology provides practical solutions to a wide range of applications, such as personalized access to local sport events through a web portal, cost-effective and fully automated production of content dedicated to small-audience, or even automatic log in of annotations.**

***Index Terms*—Production of Personalized Video Summarization, Content Repurposing, Multi-sensored Processing**

I. Introduction

Today's media consumption evolves towards increased user-centric adaptation of contents, to meet the requirements of users having different expectations in terms of story-telling and heterogeneous constraints in terms of access devices. Individuals and organizations want to access dedicated contents through a personalized service that is able to provide what they are interested in, at the time when they want it and through the distribution channel of their choice.

To address such kind of demands, this paper presents a unified framework for cost-effective and autonomous generation and distribution of contents from multi-sensored data, with an emphasis on team sports use cases. It first investigates the automatic extraction of intelligent contents from a network of sensors distributed around the scene at hand. Here, intelligence refers to the identification of salient segments within the audiovisual content, using distributed scene analysis algorithms. Second, it explains how that knowledge can be exploited to automate the production and personalize the summarization of raw video contents. In the basket-ball scenario envisioned to demonstrate our framework, the salient segments in the raw video content are identified based on player movement analysis and scoreboard monitoring. Player detection and tracking methods rely on the fusion of the foreground likelihood information computed in each view [1], [2], which allows overcoming the traditional hurdles associated to single view analysis, such as occlusions, shadows and changing illumination. Scoreboard monitoring provides valuable additional inputs to recognize the main actions of the game.

To produce semantically meaningful and perceptually comfortable video summaries based on the extraction of sub-images from the raw content, we formulate the selection of temporal segments and corresponding viewpoints in the edited summary as two independent problems, namely video production (for camerawork planning) and video summarization (for temporal content reorganization).

Although good production strategy and story organization are relative to a person's perspective, there are certain general principles whose implementation results in improved understanding of the scene, with a more enjoyable viewing experience. Especially, we identify the following three major factors to abstract the semantic and narrative requirement of video contents, i.e.,

- **Completeness**, which stands for both the integrity of view rendering in camera/viewpoint selection, and that of story-telling in summarization. A viewpoint of high completeness includes more salient objects, while a story of high completeness consists of more key actions.
- **Fineness**, which refers to the amount of details provided about the rendered action. Spatially, it favors close views. Temporally, it implies redundant story-telling, including replays. Increasing the fineness of a video does not only improve the viewing experience, but is also essential in guiding the emotional involvement of viewers by close-up shots.
- **Smoothness**, which refers to the graceful displacement of the virtual camera viewpoint, and to the continuous story-telling resulting from the selection of contiguous temporal segments. Preserving smoothness is important to avoid distracting the viewer from the story by abrupt changes of viewpoints or constant temporal jumps [3].

E-mail:chen-fan@jaist.ac.jp, damien.delannay@uclouvain.be, and christophe.devleeschouwer@uclouvain.be. This work has been partly funded by the FP7 European project APIDIS, by the WIST2 WALCOMO and WIST3 SPORTIC Walloon Region projects, and by the Belgian NSF.

Obviously, those three concepts have to be maximized to produce a meaningful and visually pleasant content. In practice however, maximization of the three concepts often results in antagonist decisions, under some limited resource constraints, typically expressed in terms of the spatial resolution and temporal duration of the produced content. For example, at fixed output video resolution, increasing completeness generally induces larger viewpoints, which in turns decreases fineness of salient objects. Similarly, increased smoothness of viewpoint movement prevents accurate pursuit of actions of interest along the time. The same observations hold regarding the selection of segments and the organization of stories along the time, under some global duration constraints.

Hence, our production/summarization system turns to search for a good balance between the three major factors. It first defines quantitative metrics to reflect completeness, fineness, and smoothness. It then formulates constrained optimization problems to balance those concepts. Both the metrics and the problem can be formulated as a function of individual user preferences, typically expressed in terms of output video resolution, or preferred camera or players actions, so that it becomes possible to personalize the produced content.

The federating objective of this paper is to present an integrated architecture for autonomous and cost-efficient production and distribution of personalized video summaries. To achieve this objective, it surveys/reviews some of our earlier works[1][4][5][6], but also introduces a bunch of novel contributions. Specifically, regarding the autonomous camerawork planning, and in addition to our previous work described in [4], we propose a new criterion, which has a clearer physical meaning and several calibration guidelines for efficient parameter determination, to drive the selection of the so-called optimal viewpoints to render a team-sport scene, along with a computationally efficient searching algorithm. Regarding summarization, we propose to extend our resource allocation formulation [5][6] to schedule the streaming of a concatenated subset of automatically produced and pre-encoded clips. This is an improvement compared to [5][6], which only considers the summarization of manually produced content, and relies on the detection of manually edited video shots to organize the summariy.

The paper is organized as follows. Section II surveys the previous related works, both in terms of autonomous camerawork planning and video summarization. Section III presents an overview of our integrated framework for personalized content distribution, with a short presentation of how personalization of the streamed content can be achieved at low computation cost, simply based on the concatenation of pre-encoded video clips. The paper then focuses on how to generate those clips automatically, and how to control their manipulation within the server, to forward personalized contents, adapted to user constraints or preferences. Section IV explains how multi-view analysis does support content generation. It first surveys our solution for players detection and tracking, as required by autonomous production tools. It then presents how those data are completed by the scoreboard information to recognize the main actions of a basket-ball game, so as to support personalized summarization. Section V then presents our integrated framework for autonomous production of personalized summaries. Section VI demonstrates the effectiveness of the approach, whilst Section VII concludes.

II. Related Works

Regarding the camerawork planning in autonomous production, we interpret the planning of ”virtual” camera actions as selecting a camera view and its in-frame viewpoint, rather than synthesizing a free-viewpoint scene. The related previous works are roughly classified into three major categories:

- Event-triggered selection: Camera switching or viewpoint movements are triggered by certain activities detected in the scene from audiovisual clues, such as an object entering the field of view or an audio event happening. [7] and [8] consider a meeting room scenario, and switch to the camera that displays the speaker. Event-triggered systems usually target at people-sparse and low-activity scenarios, and perform selection based on naive but explicit rules.
- Rule-based selection: More complicated conditions of camera switching can be achieved by introducing semantic or cinematic rules, relying on the analysis of objects, events and other contextual information. [7] used decaying curves to avoid fast camera switching and suppress too long shots in multimodal meetings. [9] and [10] selected a best shot from a list of candidate shots of each scene for a video conference or a multiplayer game TV show, according to pre-defined cinematic rules. [11] studied camera selection for athletic videos based on rules explicitly defined on user preferences and the characteristics of athletic events. The most challenge task is to extract explicit rules based on the integrated knowledge derived from scene understanding algorithms. For conference or athletic videos, it is possible to identify the dominant object of the scene, such as the speaker or the leading runner. Following this dominant object provides a reasonable and effective base to those rules. However, it is difficult to guide all camera/viewpoint selection with pre-defined rules for people-dense scenarios, such as basketball, where players change their speeds and directions all the time and the ball is passing rapidly between players.
- Data-driven selection: Rather than defining explicit rules, methods in this category adaptively adjust camera and viewpoints by evaluating some criteria defined on the current contextual configuration. There are some methods proposed in the literature for selecting the most representative area from a standalone image [12][13], based on some visual attention model [14]. In contrast, we presented an automatic video production system in [4], where the optimal camera/viewpoint is found by evaluating some global metrics about the completeness, fineness and occlusion of the scene, under the specified user preference. Compared to event-triggered or rule-based methods, data-driven selection is able to deal with people-dense, high activity scenarios, such as team-sports, in a flexible and efficient manner.

Summarization implies selection of temporal segments and local stories organization. Here, we identify two classes of automatic methods that have addressed this problem in previous literature:

- Methods targeting clustering of visual stimuli: Many works interpreted video summarization as extracting a short video sequence of a desired length from native video content, 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 cluster similar frames/shots into so called key frames [15][16], or solve constrained optimization of objective functions [17][18]. Since they attempt to preserve as much as possible of the initial content, all those methods are well suited to support efficient browsing applications.
- Methods targeting story-telling and semantic relevance: End-users motivation in viewing summaries is not limited to fast browsing of all clips in the whole video content. It also includes the intention to enjoy a concise video with well-organized story-telling and retrieval of semantically meaningful events that best satisfy users interest. Regarding semantic relevance, we observe that many works have been devoted to the automatic detection of key actions in sport events, especially for football games [19][20][21][22][23][24]. However, when addressing the problem of summary organization from actions, all those methods just implement pre-defined filtering or ranking procedures to extract the actions of interest from the original audiovisual stream. Typically, 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 [24], without taking care of story-telling artifacts. In contrast, [25] considers the continuity of the clips included in the generated summary to improve story-telling, and [26] organizes stories by considering a graph model for managing semantic relations among concept entities. Compared to general videos, stories in sport videos have much simpler structures and a limited set of possible events, which allows for both local and global control of story-telling without the need for sophisticated ontology or semantic graph models, as demonstrated by our work [5] in the context of soccer summarization. It unifies all previous works, in the sense of exploiting all kind of available knowledge, related to either production principles or the semantic of events. It goes beyond previous works by offering a flexible and generic resource allocation framework to adaptively select audio-visual segments into the summary according to user preferences. By evaluating the benefit of segments from both the content and the presentation style of the summary, our framework is able to balance the semantic (what is included in the summary) and narrative (how it is presented to the user) aspects of the summary in a natural and personal way, which is the fundamental difference of our method to filtering based approaches.

III. Personalized Content Distribution Framework Overview

We aim at designing an autonomous production system that generates and distributes personalized contents according to individual user preferences. In the present paper, we organize a summarized story through clip selection, and simulate the real camerawork by selecting a so-called optimal viewpoint (defined as a proper camera and a rectangle area within this camera view).

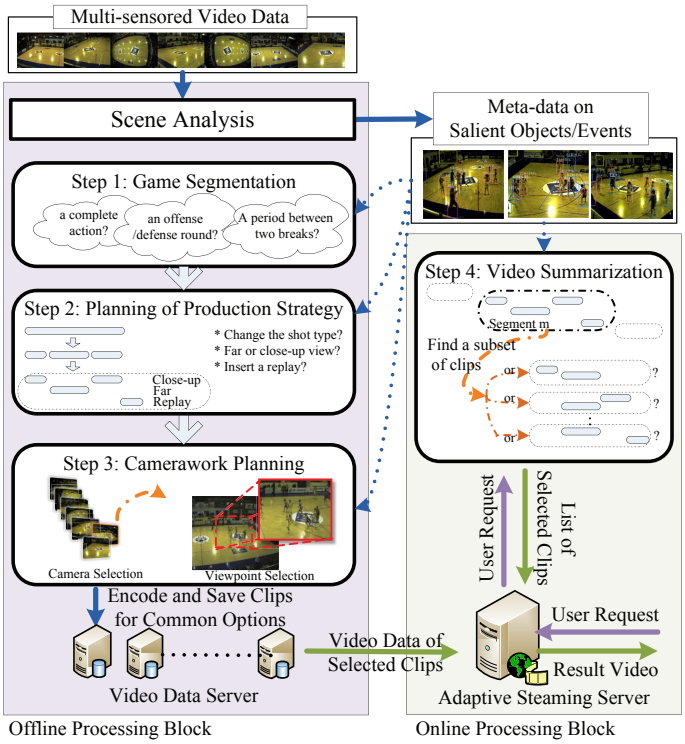


Fig. 1. The integrated framework we proposed for cost-efficient production and distribution of personalized video summaries. Besides those arrows with explicit labels, we use solid blue arrows for inputting video data, and dotted blue arrows for inputting meta-data. Wide arrows present the order of different processing steps.

We are targeting at an efficient distribution system suitable for large scale deployment, by clearly separating pre-computable parts from those that have to be handled online. Especially, by summarizing the game based on the concatenation of a subset of those clips that have been pre-encoded for several pre-specified camerawork options, we avoid computationally expensive online re-encoding of contents. In addition, according to the intrinsic hierarchical structure of video contents, both production and summarization are envisioned in the divide and conquer paradigm for improved computational efficiency. In our hierarchical video structure, a segment is defined as a video portion for a semantically complete action, which is rendered by a sequence of shots with different view-types, e.g. far view or close-up. Hence, camerawork is planned on the shot level for a continuous viewpoint sequence. A clip is the elemental component in summarization, which is obtained by further subdividing a shot. In Fig.1, we depict the overall architecture of our proposed production and distribution system, and highlight the offline and online processing stages.

From source video data, scene analysis is performed to identify salient objects and detect highlighted events. Based on these contextual metadata, we divide the whole game into short segments, where each segment covers a semantically self-contained period, e.g., an offense/defense round, a complete action, a period between two breaks, etc. For each segment, we plan the production strategy by determining the temporal boundary and view-type of each shot, and by inserting necessary replays, based on a pre-defined rule which is explained later with Fig.5, considering conventional production principles. We then perform camerawork planning for each shot that have been identified when planning the production strategy, under various options. Produced shots are further divided into short clips for finer story reorganization. The resulting edited video clips are compressed, e.g. using the H.264/AVC encoder, and stored on a disk. This ends up the offline part of the whole process, performed only once, during or just after the game.

In the online phase, given an end-user request, video summarization decides about which clips should be part of the summary, according to user preferences. The pre-encoded streams corresponding to those clips are then concatenated to be forwarded as a continuous stream to the client. The detailed description of the implementation of a streaming server supporting client-transparent concatenation of streams on a server side is beyond the scope of this paper. Interested readers can for example refer to the adaptive streaming server we have described in [28].

IV. Scene analysis

Automatic collection of contextual meta-data related to players and events is a key factor to the practicality of the above framework in a real applicative scenario. To make the whole paper self-contained, we intend to briefly introduce some computer vision technologies developed in the APIDIS project [35], which exploit multi-view analysis to support team sport actions monitoring and understanding. We first surveys our solution for players detection and tracking, which are mainly used in camerawork planning. We then present how those data are completed by the scoreboard information to recognize the main actions of a basket-ball game, so as to support personalized summarization.

A. Multi-view players detection, recognition, and tracking

Tracking multiple people in cluttered and crowded scenes is a challenging task, primarily due to occlusion between people. The problem has been extensively studied, mainly because it is common to numerous applications, ranging from (sport) event reporting to surveillance in public space. Detailed reviews of tracking research in monocular or multi-view contexts are for example provided in Yilmaz et al. [29], Khan and Shah [30][31] or Fleuret et al. [2]. In the context of team sport event monitoring, all players have similar appearance. For this reason, we focus on a particular subset of methods that do not use color models or shape cues of individual people, but instead rely on the distinction of foreground from background in each individual view to infer the ground plane locations that are occupied by people. Those methods are reviewed in Delannay et al.[1].

Fig.2 summarizes our proposed method. Similar to [2][30], our approach computes foreground likelihood independently on each view, using standard background modeling techniques. It then fuses those likelihoods by projecting them on the ground plane, thereby defining a set of so-called ground occupancy masks. The originality of our method compared to previous art is twofold. First, it computes the ground occupancy mask in a computationally efficient way, based on the implementation of integral image techniques on a well-chosen transformed version of the foreground silhouettes. Second, it proposes an original and simple greedy heuristic to handle occlusions, and alleviate the false detections occurring at the intersection of the masks projected from distinct players silhouettes by distinct views. In final, our method appears to improve the state of the art both in terms of computational efficiency and detection reliability, reducing the error rate by one order of magnitude, typically from 10 to 1%. Due to the lack of space, we encourage the interested reader to access the description presented in [1] for more details. Once players and referee have been localized, the system has to decide who's who. Therefore, histogram analysis is performed on the expected body area of each detected person. Histogram peaks extraction allows to assign a team label to each detection (see bounding boxes color in Fig.2). Further segmentation and analysis of the regions composing the expected body area permits to detect and recognize the digit(s) printed on the players shirts when they face the camera [1].

Since the player digit can only be read when the players back faces one of the cameras, we have to trace the detected players across time. Therefore, we have implemented a rudimentary whilst effective algorithm. The tracks propagation is currently done over a 1-frame horizon, based on the Munkres general assignment algorithm [32]. Gating is used to prevent unlikely matches. A high level analysis module is also used to link together partial tracks based on shirt color and/or player digit estimation. Such post-linking process only occurs when it is non-ambiguous, i.e. when one option is much more likely compared to other linking options.

B. Multi-view ball detection

The proposed approach to detect the ball relies on a set of binary masks, each mask being computed independently in each view, so as to discriminate the ball pixels from rest of the scene. In our case, the mask relies on background subtraction, completed by a cleanup pass that removes small, large, non-circular or non-moving connected regions. The ball is then expected to appear as a circular silhouette, when not occluded by players. When relevant, any additional knowledge about the texture or the color of the ball could easily be exploited. Given the masks in all views, we consider two detection strategies. The first one investigates whether a pre-defined set of 3D positions, selected along a regular 3D grid, are likely to support the ball or not. This is implemented efficiently by using integral images to approximate the correlation of the masks with the projected ball template. In contrast, the second approach selects 2D candidates in each individual view, and then computes 3D ball positions based on the

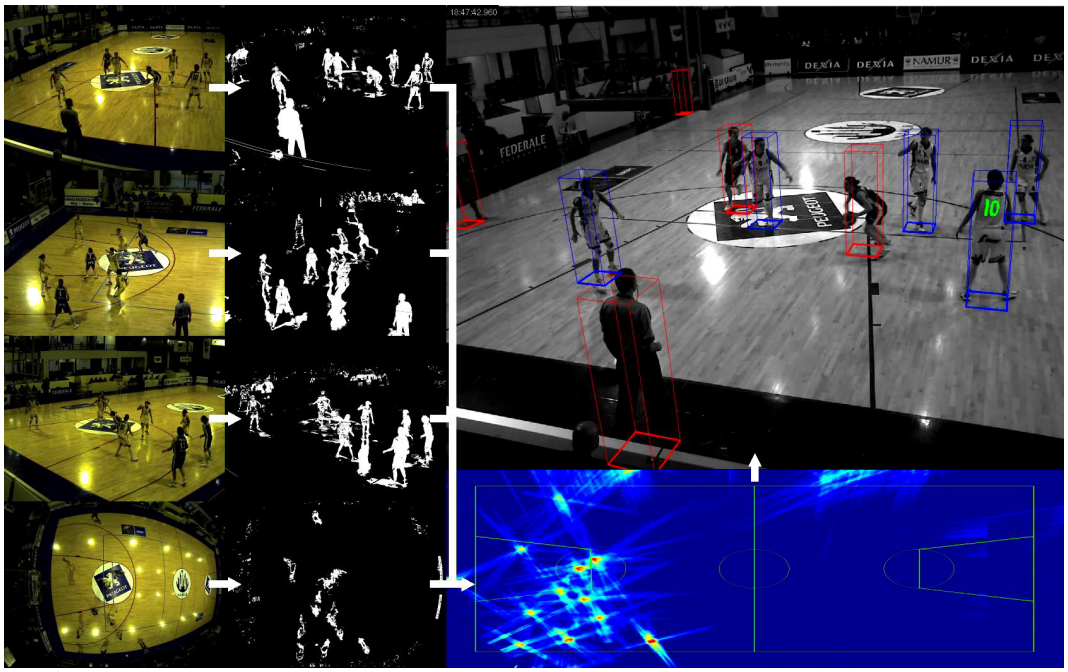


Fig. 2. Player detection and recognition: On the left, the foreground likelihoods are extracted in each view. They are projected and accumulated to the homography ground plane to define a ground occupancy map (bottom right), from which people positions are extracted through an occlusion-aware greedy process. The digits lying on the players shirts that face the camera are then recognized (top right).

triangulation and validation of those 2D candidates. Once plausible 3D ball candidates have been detected, the analysis of the candidates trajectory helps in discriminating between true and false positives, since the ball is supposed to follow a ballistic trajectory, which is not the case of most of the false detections (corresponding to body parts). Experimental results are presented in [27], which demonstrates the effectiveness and complementarity of both approaches.

C. Event recognition

This section summarizes how to detect and recognize the main actions occurring during a basketball game, i.e. field goals, violations, fouls, balls out-of-bounds, free-throws, throw-in (ball back to court), throw, rebounds, and lost balls. All those actions correspond to clock-events, i.e. they cause a stop, start or re-initialization of the 24" clock, or do occur during periods for which the clock is stopped.

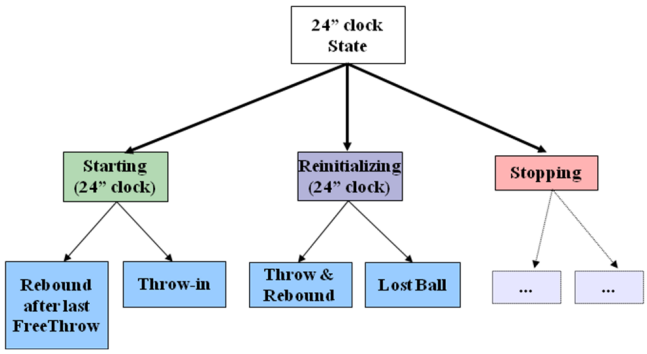


Fig. 3. Basket-ball action tree structure.

Hence, by assuming an accurate monitoring of the 24" clock and of the scoreboard, we propose to organize the actions hierarchically, as a function of the observed clock and scoreboard status. This results in the tree structure depicted in Figs.3 and 4. Most of the tests implemented in the nodes of the tree only rely on the clock and scoreboard information. When needed, this information is completed by visual hints, typically provided as outcomes of the players (and ball) tracking algorithms. The initial instance of our system defines dedicated if-then-else rules to decide about the branch to go in each node. As an example, the decision to take after a start of the 24" clock - on the left node of Fig.3 - about a "rebound" or "throw-in" action can be inferred from the analysis of the trajectories of the players. A detailed description of the detectors involved in the nodes of this tree is beyond the scope of this paper, and can be accessed in Devaux et al. [33]. Experiments demonstrate that the approach achieves above 90% accuracy.

V. Autonomous production of personalized summaries

With the contextual meta-data, we know when a user's favorite event happens or where his favorite player stands. However, practical content adaption usually requires more than simply filtering those corresponding events/objects, due to the global constraints and also possible conflicts between different user preferences. Furthermore, in order to produce visually comfortable video contents, extra regulation constraints have to be included into the production process to suppress visual/story-telling artifacts caused by sudden camera switching or incomplete story. In the following sections, we explain our solution following the four steps given in Fig.1.

A. Clock-event based game segmentation

Game segmentation enables local processing of video stories in our divide-and-conquer paradigm for more efficient

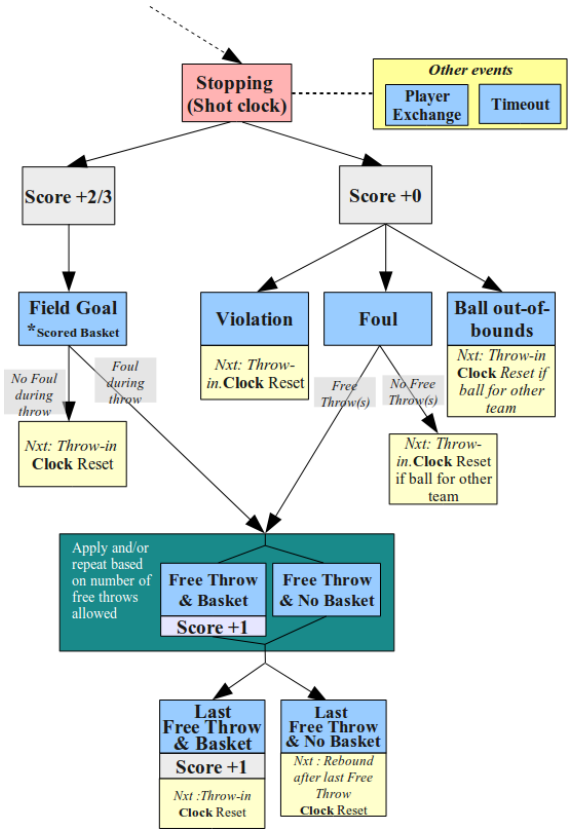


Fig. 4. Basket-ball action tree structure (Continue).

production, which could be implemented in various ways depending on the different characteristic of the target team sports, e.g, we could naturally segment a volleyball game into rallies and a baseball game according to pitchings. In basketball, the attacking team needs to attempt a shot within 24" of gaining possession, according to the 24 seconds rule. Many key events, including shooting, foul, interception and others, are closely related to starting/ending/restarting of clock counting, which are named "Start", "End" and "Reinit" clock event, respectively. We thus use clock events as a reasonable base for basketball game segmentation. Since most of critical actions (e.g. successful shots or fouls) lead to "End" clock event, it is safe and better to include all "Reinit" events in the same segment for a complete local story. More specifically, we define a segment as the period between two consecutive "Start" clocking events, as shown in Fig.5, where the clocking information could be obtained by either easily recording from the clocking system or by analyzing the score board.

B. Planning of production strategies

Shots of various view types play an important role in telling an attractive story: far views are used to present the complexity of team sports, while close-up views are essential to increase the emotional involvement of audiences. For more important actions, replays should be appended for clearer explanation of local details [3]. In conventional sport video production, a director always has a rough advance plan in his mind on how to organize those factors to present the game to viewers, based on general principles of sport video production. We

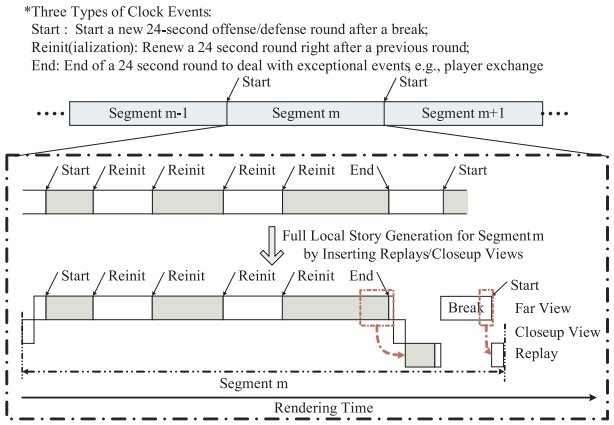


Fig. 5. Based on clock events, we divide the video into segments, and plan the full local story of each segment, including both view-type determination and replay insertion.

simulate this process, and plan the production strategy for each segment before any computationally expensive optimization in corresponding viewpoint selections, by only analyzing the structure of clock-events.

A basic rule of sport video production is to avoid dramatic camera/shot-type switching when critical action is taking place in the playing area [3]. Close-up views, replays and commercials are thus usually inserted during game breaks. However, compared to other team sports such as volleyball and soccer, basketball game runs in a much faster pace, due to the fact that the 24" shot clock could be reinitialized right after the previous round. By exploring the view structure of three broadcasted sport videos [6], we confirm that a basketball video is rendered in far/medium views for most of the time, because the director has only a few chances to insert close-up views and replays. We plan the production strategy by following this production convention. Furthermore, since it usually takes the audiences about one second to reestablish their relationship to the game after camera/view switching [3], we prefer to include a temporal bias for a smoother transition between view types.

As shown in Fig.5, we take the following reference strategy to render each segment, i.e.,

(1) A 2" close-up is taken 1" $\sim$ 2" before the "Start" event.

(2) A 1" $\sim$ 2" close-up is inserted after the "End" event, depending on the length of break to the next "start" event.

(3) A replay is inserted to cover a period from the last 1/3 part of the round before the "End" event to the starting point of the close-up in (2), if this period is longer than 3".

(4) We also insert a replay for the break between the end event and the next start event. Note that this break overlaps with the next segment, which we think is necessary to help audiences to reorient themselves in the new segment.

(5) Other parts of the segment are rendered as far views.

C. Autonomous production

According to the view-type planned, we start to implement the camerawork in a personalized way that reflects users' preferences and device limitations. Three kinds of camera movement could be simulated: camera rotation is able to be

partially implemented via camera switching, while camera zooming and translation are approximated by cropping.

In our first trial in [4], we built a viewpoint selector by picking a Mexican-hat like function to trade off between completeness (including more objects) and fineness (enlarging screen sizes of objects), based on manually annotated objects independently collected from each camera view. As a specially designed function without a (real or assumed) physical explanation, it is neither easy to select proper parameters for this criterion nor simple to further improve it. This criterion also fails to reflect the correct relationship between different camera views, which thus forced us to define another criterion for camera selection in [4], where inconsistency between these two criteria was left as unsolved.

Automatically detected salient objects in Section IV-A preserve inter-camera relationship in two aspects: 1.) At each moment, all camera views share the same universal set of salient objects, although some objects are only visible in a subset of camera views. Accordingly, the completeness of a viewpoint could be directly evaluated from the number of its visible objects; 2.) The bounding-boxes obtained by re-projecting the identical 3D salient object into all camera views partially reflect the relative camera positions, such as close/far or view angle to the ground plane. By exploiting those information, we are allowed to design a unified criterion for both camera selection and viewpoint selection.

Basically, we optimize the completeness by maximizing the accumulated interests of various salient objects, as in [4]. This results in the tendency to enlarge the viewpoint, which is withheld from the request to maximize the fineness. In this paper, we reformulate the criterion, where the role of each term in completeness, fineness and occlusion is clearer. We further summarize the following guidelines of camera/viewpoint selection, which serve as calibration baselines to refine the relative significance of different factors in the proposed criterion so as to stabilize its performance.

- In the ideal case where all salient objects have the same importance and are evenly distributed in the scene, the size of the optimal viewpoint should be equal to the specified display resolution so as to avoid quality loss due to image resampling;
- For two viewpoints of different sizes which have completely the same coverage of the play court, the defined criterion should output the same benefit value since they are identical when rendered using the display resolution;
- In team sport broadcasting, most of the scenes are provided by side cameras, which is not only due to their lower installation costs (especially for outdoor games) but also because they could present player's local activity as well as the global team work.[1] It also looks more natural to the audiences that players stand vertically and move horizontally. From two camera positions that have the same completeness and occlusion, we prefer the closer one if they have the same view angle to the ground plane, and prefer the one of a smaller view angle if they have the same distance.

[1]Similarly, in a swimming match, more scenes come from ceiling cameras and underwater cameras.

Formally, we define a viewpoint $\mathbf{v}_{ki}$ in the k^{th} camera view of the i^{th} frame, by its size S_{ki} and its center $\mathbf{c}_{ki}$. Assume that N salient objects have been detected in a frame, and the location of the n^{th} object in the k^{th} view is denoted by $\mathbf{x}_{nki}$. Owing to player identification, we define the importance of the n-th object as a function $I_n(u^P, u^T)$ of a preferred player u^P or a preferred team u^T. For producing far views and replays, we select the optimal camera k^* and its viewpoint $\mathbf{v}^*_{ki}$, by maximizing the overall benefit

$$\mathcal{B}_{ki}(\mathbf{v}_{ki}|\mathbf{u}) \quad = \quad w_k(u^C)\sum_{n=1}^{N} w_{ki}(\mathbf{x}_{nki})I_n(u^P, u^T), \quad (1)$$

where $w_k(u^C)$ denotes the weight assigned to the k^{th} camera, so as to force the system to favor a user-preferred camera u^C. The attentional significance of each salient object within the present viewpoint is weighted by

$$w_{ki}(\mathbf{x}_{nki}) \quad = \quad \alpha(.)\beta(.)o_k(\mathbf{x}_{nki}). \quad (2)$$

In the above equations:

a) Function $\alpha(.)$ modulates the weights of the objects according to their distance to the scene center, normalized by the viewpoint size. This weight should be high and positive when the object-of-interest is within the viewpoint and close to the scene center $\mathbf{c}^{\mathrm{SCN}}_{ki}$, and should be negative or zero when the object lies outside the viewing area. Especially, we use the following $\alpha(.)$, i.e.,

$$\alpha(.) = \exp\left(-\frac{\left\|\mathbf{x}_{nki} - \mathbf{c}^{\mathrm{SCN}}_{ki}\right\|^2}{2S^2_{ki}}\right) \times \mathcal{V}(\mathbf{x}_{nki}, S_{ki}, \mathbf{c}_{ki}), \quad (3)$$

where visibility function $\mathcal{V}(\mathbf{x}_{nki}, S_{ki}, \mathbf{c}_{ki})$ takes 1 if object $\mathbf{x}_{nki}$ is fully covered by viewpoint $\mathbf{v}_{ki}$, and 0 for not. We set $\mathbf{c}^{\mathrm{SCN}}_{ki}$ to the ball position (or the gravity center of all objects when ball position is not available) so as to cover more key objects in the current action, based on a simple assumption that dominant players usually surround the ball.

b) Function $\beta(.)$ reflects the penalty induced when the native signal of the k-th camera has to be sub-sampled once the viewpoint size becomes larger than the maximal resolution u^{res} allowed by the user. An appropriate choice consists in setting the function equal to one when $S_{ki} < u^{res}$, and in making it decrease afterwards, e.g.,

$$\beta(.) = h_k(\mathbf{x}_{nki})\left[\min\left(\frac{u^{res}}{S_{ki}}, 1\right)\right]^{u^{close}}, \quad (4)$$

where $h_k(\mathbf{x}_{nki})$ is the height in pixels of projecting a six feet tall vertical object (average height of a player) located in $\mathbf{x}_{nki}$ into camera view k, which serves as normalization of different camera views, and is directly computed based on camera calibration. Maximizing $h_k(\mathbf{x}_{nki})$ leads to either a closer camera view to the player or a smaller view angle to the ground plane, both result in higher fineness. $u^{close} > 1$ increases to favor close viewpoints compared to large zoom-out views. It assures that the loss of fineness increases faster than the benefit of completeness through further zoom-out, when the viewpoint is already larger than u^{res}.

c) $o_k(\mathbf{x}_{nki})$ measures the occlusion ratio of the n-th object in camera view k, which is defined as the fraction of pixels of the object NOT overlapped by other objects when projected to the camera view. The z-depth order could be approximately computed from the size of the bounding box $h_k(\mathbf{x}_{nki})$. For replay, we use $o_k^2(\mathbf{x}_{nki})$ instead of $o_k(\mathbf{x}_{nki})$ to emphasize more on reducing occlusions.

For conciseness, we use vector $\mathbf{u}$ to group all these user preferences mentioned above, i.e. $\mathbf{u} = [u^{close} u^{res} u^P u^T u^C]$. In the section of experiments, we will further clarify the role of each term, and also verify that the above criterion could satisfy all those guidelines we set.

Since the optimal viewpoint to maximize the benefit is not directly solvable, a search over candidate viewpoints needs to be executed. It is natural to ask the viewpoint of a far view or replay to include multiple players. In this situation, we are able to improve the efficiency of this search, by reducing the size of the solution space. Eq.(1) reflects the trade-off between fineness and completeness. If we decay the $\alpha(.)$ slow enough, reducing fineness through virtual zooming out is only beneficial if it includes additional players in the viewpoint. In other words, it is useless to enlarge the viewpoint (thereby reducing fineness) without including any additional player. As a consequence, an optimal viewpoint (of a fixed aspect ratio) will always be spanned by two players. We can thus restrict our initial full-search analysis to a selective search focusing on viewpoints of the required aspect ratio that include a pair of support players on their border. Even an exhaustive selection of all possible pairs of (maximum 10) players can reduce computational complexity dramatically, compared to the grid search used in [4].

We assign higher $w_k(u^C)$ to two major side-view cameras for far view, while for replays other cameras have higher $w_k(u^C)$, so as to produce different results for normal game plays and replays. For close-up views, we simply find the minimal box that covers the closest player to the scene center $\mathbf{c}_{ki}^{\mathrm{SCN}}$. [2] When the viewpoint is determined, we expand the viewpoint by 10% in all directions to leave a margin space for better appearance, as in conventional production[3].

Finally, an iterative smoothing process based on a two-layer Markov chain is applied to the selected sequences of viewpoints to remove visual artifacts such as flicking and fluctuation of the view (see [4]). In the first layer, a camera-wise smoothing is committed to stabilize the viewpoint motion in each camera, based on the sequence of optimal viewpoints obtained in each individual camera view. Camera benefits, which are evaluated on each individual frame based on the determined viewpoints, are then fed into the second layer of the smoothing process as the a priori knowledge so as to recover a smooth camera sequence without abrupt camera switching.

After generating the viewpoint sequences, we cut long shots into shorter clips ($\sim$ 2"). It is worth mentioning here that the reference rendering strategy does not define the actual way a segment is rendered. We have only constructed a universal set of pre-encoded clips. Any subset of the clips of a segment actually defines an eligible rendering strategy. An optimal strategy is then selected by the summarization method described in the next section.

[2]Obviously it makes more sense to zoom on the player of interest (e.g. preferred player, or player who scored). We provide this as a provisional implementation before we have reliable automatic player identification, so as to complete the overall framework and leave space for future local revisions.

D. Personalized summarization

A summary is organized by proper selection of clips, so as to satisfy both semantic preferences in terms of action or player and narrative preferences in terms of story-telling patterns (replays or not, long or short segment stories). This process is based on the generic resource allocation framework, which has been verified to be efficient in summarizing broadcasted sport videos [5]. Fig.6 briefly reminds the proposed summarization framework. As explained in Section V-A, 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. 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.

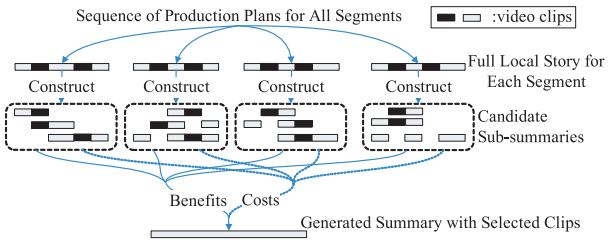


Fig. 6. A resource allocation based framework of sport video summarization.

Formally, for the m-th segment $\mathbf{S}_m$, we consider L different narrative options $\{\mathbf{s}_{ml}\}$, each option defining the subset of the clips of this segment that are rendered during display. A pair of benefit/cost values, i.e., $\mathcal{B}(\mathbf{s}_{ml})$ and $\mathcal{C}(\mathbf{s}_{ml})$, is assigned to each option $\mathbf{s}_m$, and a summary is obtained by maximizing the overall benefit under the length constraint u^{LEN}, i.e.,

$$\{\mathbf{s}_{ml}^*\} = \arg\max_{\{\mathbf{s}_{ml}\}} \sum_m \mathcal{B}(\mathbf{s}_{ml}), \quad \sum_m \mathcal{C}(\mathbf{s}_{ml}) = u^{LEN}, \qquad (5)$$

which is thus able to be modeled as a resource allocation problem. 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 (e.g., $\sum_m \mathcal{C}(\mathbf{s}_{ml}) \leq u^{LEN}$), 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[34].

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. The main theorem of applying the Lagrangian relaxation to this convex-hull

approximated resource allocation problem reads that if λ is a non-negative Lagrangian multiplier and $\{\mathbf{s}^*_{ml}\}$ is the optimal set that maximizes

$$\mathcal{L}(\{\mathbf{s}_{ml}\}) = \sum_m \mathcal{B}(\mathbf{s}_{ml}) - \lambda \sum_m \mathcal{C}(\mathbf{s}_{ml}) \quad (6)$$

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

$$\sum_m \mathcal{B}(\mathbf{s}_{ml}) - \lambda \sum_m \mathcal{C}(\mathbf{s}_{ml}) = \sum_m (\mathcal{B}(\mathbf{s}_{ml}) - \lambda \mathcal{C}(\mathbf{s}_{ml})). \quad (7)$$

From the curves of $\mathcal{B}(\mathbf{s}_{ml})$ with respect to their corresponding summary length $\mathcal{C}(\mathbf{s}_{ml})$, the collection of points maximizing $\mathcal{B}(\mathbf{s}_{ml}) - \lambda \mathcal{C}(\mathbf{s}_{ml})$ 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 $\mathcal{B}(\mathbf{s}_{ml})$ with respect to $\mathcal{C}(\mathbf{s}_{ml})$, we can use a greedy algorithm to search for the optimum under a given constraint u^{LEN}. 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 then define the sub-summaries for each segment.

One major advantage of using this summarization method is that it allows highly personalized nonlinear story organization via flexible definition of benefits. In the present paper, the benefit is defined as

$$\mathcal{B}(\mathbf{s}_{ml}) = \sum_{j \in \mathbf{s}_{ml}} \mathcal{I}_{mlj} \mathcal{G}(\mathbf{s}_{ml}, u^P, u^T) \mathcal{P}^{CR}_{ml}(u^C, u^R) \mathcal{P}^F_{ml}, \quad (8)$$

which includes accumulated semantic importance of selected clips $\sum_{j \in \mathbf{s}_{ml}} \mathcal{I}_{mlj}$ and extra gain $\mathcal{G}(\mathbf{s}_{ml}, u^P, u^T)$ from user favorite player u^P and team u^T, and also evaluates narrative preferences on story-telling (e.g., penalty $\mathcal{P}^{CR}_{mj}$ on user specified story continuity u^C, and story redundancy u^R). Satisfaction of general production principles is also evaluated through the penalty for forbidden cases $\mathcal{P}^F_{mj}$, to avoid frustrating visual/story-telling artifacts(e.g., over-short/incomplete local stories).

VI. Experimental validation

We organized a data-acquisition in the city of Namur, Belgium, under real game environment, where seven cameras were used to record four games. All those videos are publicly distributed in the website of APIDIS project [35] and more detailed explanation about the acquisition settings could be found in [36]. In Fig.7, sample images from these cameras are given. Performance evaluation of the proposed production system has been committed based on these videos.

Fig. 7. Sample views gathered by different cameras.

In the following part of this section, we first investigate the behavior of the proposed criterion for viewpoint selection. We then present some result summaries, followed by subjective evaluation results, to verify the capability of our system in satisfying various user preferences.

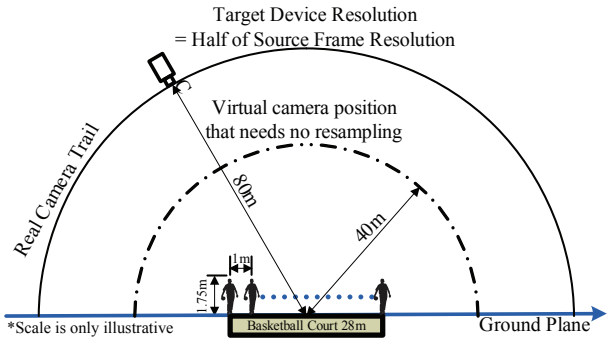
A. Behavior of the viewpoint selection criterion

Due to infinite possible configurations of player positions, it is difficult to evaluate whether the viewpoint selector fulfills the design goal, by directly inspecting the determined viewpoints from a real game video. Here we use an ideal sandbox case for verifying the behavior of the proposed criterion for both camera and viewpoint selection.

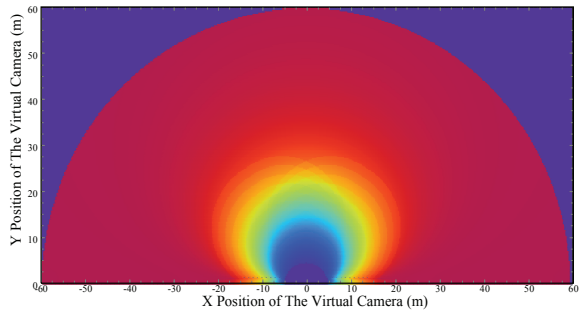
As depicted in Fig.8(a), we consider a special case where players are evenly distributed along the 28m long central line of the basketball court. Those players are of the same height 1.75m and the distance between any two consecutive players is set to 1m. Accordingly, we have 29 players in total. By moving a pinhole camera along the circular trail, we collect source camera views from all angles. The radius of this circular trail should be large enough, which is 80m here, so as to assure that the optimal viewpoint is covered by one of these camera views. Without loss of generality, we assume that we intend to find an optimal viewpoint for a target display, whose resolution is only half of that of this pinhole camera. Intuitively, we conclude that the viewpoint that needs no resampling should have the same size of the target resolution, which is equivalent to put the virtual camera 40m away.

For each virtual camera position with a distance ranging from 5m to 60m, we compute its equivalent viewpoint in the corresponding source camera view, where positions of players within this viewpoint are easily computable by using projective geometry. We then use the proposed criterion to compute the benefit of this virtual camera position[3]. We plot these benefits in Fig.8(b)-(d), by using both the complete form of the criterion and two incomplete forms with certain terms missing, which helps us to clarify the exact role of each term. When only completeness is considered (by omitting $\beta(.)$ function and the occlusion term), enlarging the viewpoint to include more players is always beneficial, which drives the virtual camera far away (Fig.8(b)). Inclusion of $\beta(.)$ function leads to three obvious changes, as revealed by Fig.8(c). 1) The tendency of enlarging the viewpoint is withheld by the will to have a larger pixel size for each player, where a trade-off has been built up; 2) A virtual camera with parallel optical axis to the ground plane is most favored, since they increase both the number and the pixel size of visible players, without considering the occlusion; 3) A circular ridge starts to appear around 40m. This ridge becomes much clearer in

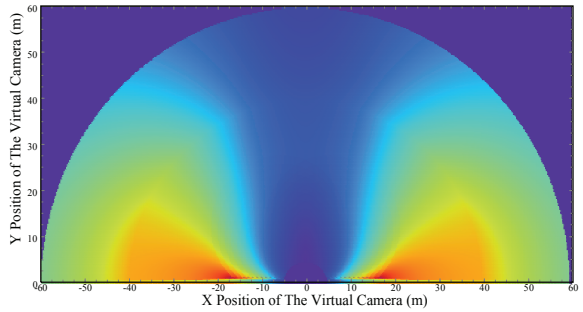
[3]Here, the camera weight $w_k(u^C)$ is always set to 1, and all players have the same interest.



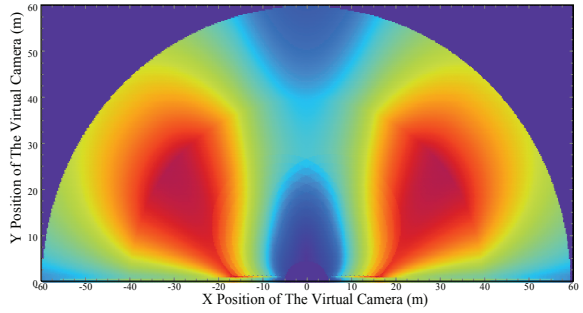
(a) A sandbox case for investigating the criterion



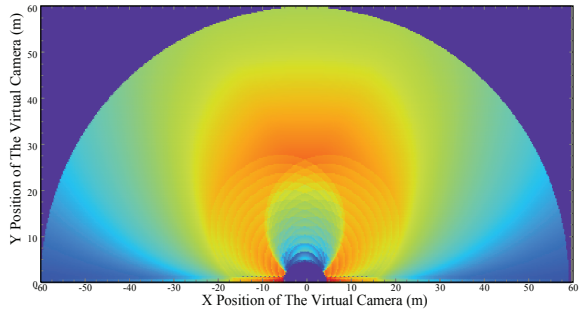
(b) Incomplete form with only completenss



(c) Incomplete form without occlusion



(d) The complete form of proposed criterion



(e) The old version we proposed in [4]

Fig. 8. Maps of benefits computed at various virtual camera positions are plotted in a sandbox ideal case, so as to clarify the role of each term in the proposed criterion and make comparisons to a previous version in [4].

Fig.8(d), where an even more balanced benefit is computed by further considering the occlusion term. In Fig.8(d), the maximal benefit is achieved from virtual cameras with an oblique view angle to the ground plane, among which those positions on the 40m circle are further favored so as to avoid unnecessary resampling, which coincides with our intuitive understanding and predefined guidelines about the optimal viewpoint. When the target resolution changes, the optimal circular ridge moves accordingly, which hence realizes the personalization against device resolutions.

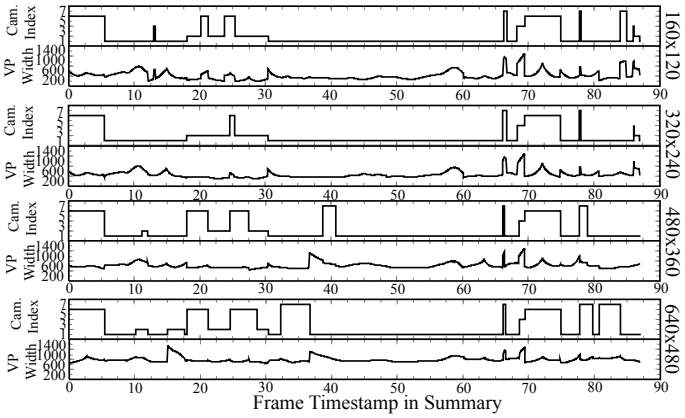


Fig. 9. Selected camera index and viewpoint sizes of a 90" long summary produced under four resolutions, i.e., 160×120, 320×240, 480×360, 640×480.

In Fig.8(e), we also plot the criterion we proposed in [4]. Although this old version also built up the balance between completeness and fineness, it is difficult to determine the best settings of parameters. Furthermore, due to lacking normalization of object size in different camera views, this criterion favors top views more, because they could include more players with less occlusion. Compared to that, the new criterion considers the normalization between different camera views and is thus able to perform more natural camera selection, owing to the availability of inter-camera correspondence relations from multi-view analysis. The new criterion has a clearer meaning for all its terms, which lights up the direction of future improvements. It also has several guidelines to calibrate relative strengths of these terms, which helps to tune parameters and makes it practically more useful.

B. Autonomous production of personalized summaries

A major advantage of this proposed framework lies in that it provides flexible personalization abilities to satisfy various user preferences. We first provide some representative results so as to give the readers an initial idea of the output, and then depend on subjective evaluation in the next section for performance validation.

In Fig.9, we investigate the personalization ability in resolution adaption by plotting the resultant camera index (in the upper part of sub-graph) and viewpoint widths (in the lower part of sub-graph) of a 90" long summary under four resolutions, i.e., 160×120, 320×240, 480×360, and 640×480. It is obvious that larger viewpoints have been selected for a larger resolution. We also observe that the viewpoint sizes are

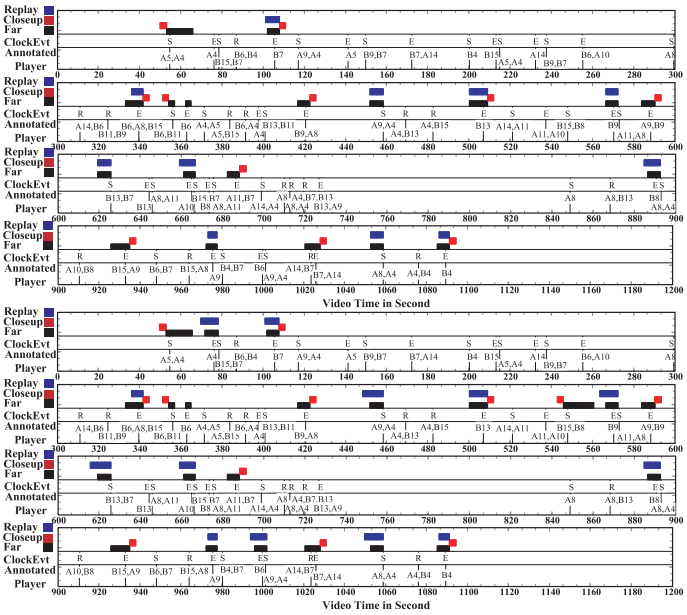


Fig. 10. Temporal ranges and view-types of selected clips by using native game time for two different length constraints, i.e., 4min(upper) and 5min(lower), along with their corresponding clocking events and dominant players.

maintained in an average level that is close to the resolution, so as to minimize the quality loss in resampling.

Fig.10 presents two results with different length constraints, namely, 4min and 5min. Within each sub-graph, we plot the temporal ranges and view-types of selected clips by using native game time, along with their clocking events and dominant players. It reveals that our summarization method does not simply expand story around pre-detected hot-spots, but intends to reallocate the time resources to render events of interests, considering both local and global story-telling.

Interested readers can access more summary samples under different user preferences at [35]. A demo system including both the production tool and all required data is also downloadable through the above webpage.

C. Subjective Evaluation

Subjective evaluations have been performed separately in [4] and [6], relating to viewpoint selection and summarization, respectively. Produced results have been evaluated from both their global impressions and visual/story-telling artifacts, where the efficiency of each corresponding method has been confirmed. The main contribution of the paper lies in integrating our video analysis and production/summarization components into a fully automatic framework. Therefore, novel subjective evaluations are performed here to address: 1.) The relevance of our personalized video production concept; 2.) The efficiency of our proposed implementation in achieving this personalized production.

Relevance of personalized summarization concept

We interviewed 17 people for their opinions about our personalized production concept. Interestingly, the panel corresponds to a representative set of users since they include 4 content production professionals, 3 sport professionals, 6 basket-ball fans, and 4 computer vision experts. In the questionnaire (downloadable from the supplemental page [35]), we considered several content creation scenarios [full game production for VoD, summary generation, interactive browsing, and access to game/players statistics], and asked them to write down what they would expect from a personalized service for each application scenario. We also prepared a list of personalization criteria, including many game actions, statistics and audio/graphic elements, and asked them to evaluate the importance of each criterion.

User's understanding/feeling of personalized summarization concept has been evaluated before and after playing with our summarization prototype, which is available in the supplemental page [35]. Hence, we summarize their feedbacks into two groups. The a priori opinions collected before playing with the prototype are summarized as follows:

- All application scenarios envisioned in the questionnaire (production, summarization, browsing, statistical analysis) are considered to be relevant to at least one kind of end-user;
- The level of interest of a given user towards a specific scenario depends on his/her professional background. Coaches and players are especially interested in statistics and browsing capabilities. Fans are interested in generic and personalized summaries. Content production professionals are primarily interested in raw content provisioning at low cost, e.g. for VoD services or to feed the manual construction of summaries.
- When considering the personalization/browsing criteria, it appears that recognition of both the action and its associated players is important. Classification of structured actions, i.e., selection of actions for which a given player receives the ball in a particular place, only interests coaches. Zoomed-in views are of little importance to sport coaches and players, and only interest content producers, if sufficient resolution can be preserved.

After having played with the prototype, the following additional conclusions could be drawn:

- The personalization criteria currently implemented in our summarization test-bed are considered as being relevant. However, their implementation effectiveness is generally evaluated to medium or even low for the replays and resolution. Based on users oral feedbacks, we conclude that this is probably due to the fact that (1) the image quality degrades a lot when zooming-in the picture, and (2) replays are not properly inserted in the narrative flow. They are played very fast, and do not always provide a different point-of-view on the action.
- Additional personalization criteria have been pointed out by the users, including the opportunity to select a time period of the game, and the period of the game during which one specific player was on the field. Sport professionals pointed out the fact that all criteria listed in the questionnaire were relevant. Augmenting a top view with the label of each player would also be useful.
- Audio support and graphics elements (score, timer, etc) are considered as a fundamental component of the summary. The absence of those components in our test-bed has often been pointed as one of its main drawbacks.

- Most of the interviewees believe that consumers would be ready to pay for such personalized services. Fans consider that the services should be part of a provider package. Content distribution professionals would be ready to pay a fee to access high-quality contents. Sports professionals (clubs, managers, coaches, etc) tolerate lower quality content, but request personalized access mechanisms.

Quantitative evaluation of user satisfaction

Besides the qualitative impression over the prototype, we also perform a standalone subjective evaluation via "mean opinion scores" to provide some quantitative results. We prepared a webpage as given in [35], which presents 5 groups of videos under different user-preferences. Viewers were asked to score the relevance of several personalization criteria, and also the effectiveness of our implementation in personalizing the video with respect to each investigated criterion. Both notions were scored using four ranks, i.e., "Very High"(=4), "High"(=3), "Low"(=2) and "Very Low"(=1). Five criteria have been selected, i.e., "Display Resolution", "Replay Insertion", "Preferred Event", "Preferred Player" and "Summary Duration". We collected answers from 20 persons, and plotted the mean scores and their standard deviations in Fig.11.

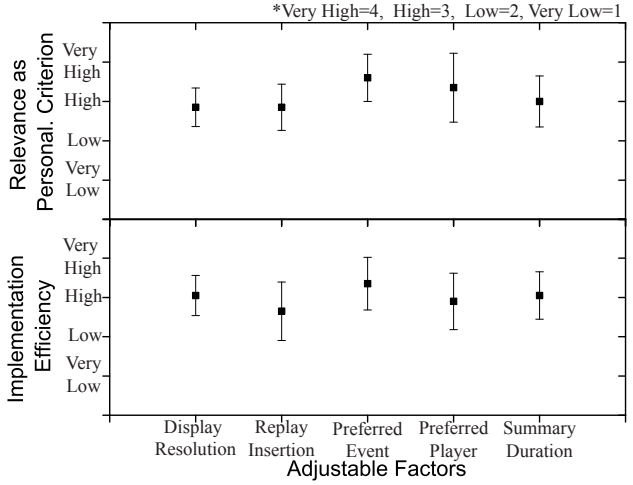


Fig. 11. Subjective evaluation results from 20 subjects show that our proposed implementation is efficient in personalizing videos with respect to several major personalization criteria.

In the top sub-graph of Fig.11, we present the relevance of the five factors as personalization criterion, while in the bottom, we show how user appreciate the effectiveness of our implementation. We make the following major observations:

- As an overall result, all the five factors are regarded as highly relevant to personalized production, and our method is regarded as efficient to personalize the video with respect to these factors.
- "Preferred Event" and "Preferred Player" are rated as two most important personalization criteria, which coincides with conventional understanding of personalized video summaries. "Display Resolution" is introduced to multi-view video production, which was less often discussed in single-view video summarization. "Replay Insertion" is an operation against conventional understanding of summarization as producing a concise video of the original source. Therefore, it is natural to find that they are less accepted. However, we still observe that these two factors are rated as "highly" relevant, which not only validates our concept of video production from multi-view data, but also confirms our argument that video summarization should be regarded as a chance to personalize the contents rather than simply filtering important events.
- As for our implementation effectiveness, the highest score is obtained by personalization against "Preferred Event". Our implementation against "Preferred Player" is also evaluated as "highly" efficient. Hence, despite the possible incompleteness and errors of information of both events and dominant players, we are still allowed to provide meaningful results, so as to partially satisfy users' semantic preferences.

  Our personalized ability against "display resolution" obtained the second highest score, which proves the efficiency of our production system, including both camera and viewpoint selection. Our implementation against "Replay Insertion" has the lowest effectiveness, which also coincides with the overall impressions of those people after playing with our prototype. In order to improve the quality of "Replay Insertion", we need to have more accurate localization of beginning and ending points of events, and consider the proper presentation, e.g., slow playback, which is left as one of our future work.

Note about computational complexity

Regarding computational complexity, the deployment of a permanent infrastructure in the spiroudome, a major basketball stadium in Charleroi, Belgium, has proven the computational feasibility of our production concept. All stages related to IP camera image decompression, to player detection, to viewpoint selection, and to image reconstruction are running in real-time on a HP server including 2 quad-core processors. The computational complexity of the resource allocation algorithms involved in the summarization process is known to be small [6], since selection and organization of local stories only requires a few hundreds of milliseconds in the summarization test-bed. One unknown remains regarding the complexity of player recognition and tracking processes, as well as regarding the ball detection algorithms. Those stages are important to complete the information collected on the scoreboard, to recognize the actions and their key players. Their real-time implementation is however beyond the scope of this paper.

All the above results demonstrate the relevance and feasibility of the APIDIS sport production concept, but also reveal a number of challenges that need to be addressed before lucrative commercial exploitation of the concept. Specifically, even if the current solution appears to fulfill the expectations of basket-ball fans for (local) game coverage, it also appears that improved image quality or more accurate game analysis solutions are required to fully satisfy the requirements of content providers and sports professionals, respectively.

VII. Conclusions

We proposed a framework for producing personalized summaries of basketball videos from multi-sensored data. The system builds on multiview video analysis to interpret the game. By taking divide-and-conquer strategy, we efficiently

solve the problem of viewpoint determination and temporal segment selection. Especially, we defined the planning rule of production strategy and flexible criteria for viewpoint selection, and implemented a real-time production system.

One major contribution of this paper is to integrate our video analysis and production/summarization components into a fully automatic framework. Furthermore, an original content distribution framework that is suitable for large scale deployment is presented, and subjective experiments are reported to validate both the semantic relevance and the implementation effectiveness of our personalization criteria.

Our method for producing personalized video summaries has four major advantages. Namely, it offers 1.) Strong personalization opportunities: Semantic clues about the events detected in the scene can easily be taken into account to adapt camerawork or story organization to the needs of the users. 2.) Improved story-telling complying with production principles: On the one hand, production cares about smooth camera movement while focusing on semantically meaningful actions. On the other hand, summarization naturally favors continuous and complete local stories. 3) Computational efficiency: We adopt a divide-and-conquer strategy and consider a hierarchical processing, from frames to segments. 4) Generic and flexible deployment capabilities: The proposed framework balances the benefits and costs of different production strategies, where benefits and other narrative options can be defined in many ways, depending on the application context.

Subjective evaluation also highlighted our future works. In a near future, we will focus on improving the insertion of replay/close-up views and the enhancement of image qualities. Towards a practical media service, insertion of audios and other supportive information, e.g., on-screen texts and graphics, and implementation of online user interaction for more flexible content personalization should also be addressed.

All these exploit a way to provide highly personalized video services to satisfy various user preferences, not only in basketball game, but also in many other application scenarios.

REFERENCES

[1] Delannay D., Danhier N., and De Vleeschouwer C., "Detection and recognition of sports (wo)men from multiple views," *ICDSC'09*, pp.1-7, 2009.

[2] Fleuret F., Berclaz J., Lengagne R., and Fua P., "Multi-camera people tracking with a probabilistic occupancy map. " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.**30**, pp.267-282, 2008.

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

[4] Chen F., and De Vleeschouwer C., "Personalized production of team sport videos from multi-sensored data under limited display resolution. " *Computer Vision and Image Understanding, Special Issue on Sensor Fusion*, vol.**114**, pp.667-680, 2010.

[5] Chen F., De Vleeschouwer C., Barrobes H.D., Escalada J.G., and Conejero D., "Automatic and personalized summarization of audio-visual soccer feeds," *ICME'10*, pp.837-842, 2010.

[6] Chen F., De Vleeschouwer C., "Formulating team-Sport video summarization as a resource allocation problem," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.**21**, pp.193-205, 2010.

[7] Kubicek R., Zak P., Zemcik P., and Herout A., "Automatic video editing for multimodal meetings, " *ICCVG'08*, pp.1-12, 2008.

[8] Rui Y., Gupta A., and Cadiz J.J., "Viewing meetings captured by an omni-directional camera," *ACM CHI'01*, pp.450-457, 2001.

[9] Vronay D., Wang S., Zhang D., and Zhang W., "Automatic video editing for real-time multi-point video conferencing, " *US Patent*, 20060251384, 2006.

[10] Vronay D., Wang S., Zhang D., and Zhang W., "Automatic video editing for real-time generation of multiplayer game show videos," *US Patent*, 20060251383, 2006.

[11] Papaoulakis N., Doulamis N., Patrikakis C., Soldatos J., Pnevmatikakis A., and Protonotarios E., "Real-time video analysis and personalized media streaming environments for large scale athletic events," *AREA'08*, pp.105-112, 2008.

[12] Suh B., Ling H., Bederson B.B., and Jacobs D.W., "Automatic thumbnail cropping and its effectiveness," *ACM UIST'03*, pp.95-104, 2003.

[13] Xie X., Liu H., Ma W.Y., Zhang H.J., "Browsing large pictures under limited display sizes," *IEEE Transaction on Multimedia*, vol.**8**, pp.707-715, 2006.

[14] Itti L., Koch C., and Niebur E., "A model of saliency-based visual attention for rapid scene analysis, " *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.**20**, pp.1254-1259, 1998.

[15] 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.

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

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

[18] Pahalawatta P.V., Zhu L., Zhai F., and Katsaggelos A.K., "Rate-distortion optimization for internet video summarization and transmission," *MMSP'05*, pp.1-4, 2005.

[19] Qian R., and Haering N., "Method for automatic extraction of semantically significant events from video, " *US6721454 (B1)*, 2004.

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

[21] Murphy N., and Smeaton A., "Audio-visual sequence analysis, " *WO2005124686 A1, UNIV DUBLIN CITY*, Publication info: IE20040412 (A1), 2005.

[22] Jung C., Kin C., Kim S.K., Lee G., Kim W.Y., and Hwang S., "Method and apparatus for summarizing sports moving picture, " *SAMSUNG ELECTRONICS CO LTD*, JP2006148932, 2006.

[23] Pan H., and Li B.X., "Summarization of soccer video content," *US Patent*,US20040017389A1, 2004.

[24] Gong Y., "Method and apparatus for personalized multimedia summarization based upon user specified theme," *NIPPON ELECTRIC CO [JP]*, US6751776 (B1), 2004.

[25] 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.

[26] 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 Transaction on Multimedia*, vol.**11**, pp.295-312, 2009.

[27] Parisot P., Amit Kumar K.C., and De Vleeschouwer C., "Multiview ball detection based on foreground masks", *ICIP'11*, (Submitted), 2011. Supplemental Data: http://thetis.tele.ucl.ac.be/Apidis/pascaline/icip2011/BallDetectionAndTracking.htm

[28] Bömcke E., and De Vleeschouwer C., "An interactive video streaming architecture for H.264/AVC compliant players," *ICME'09*, pp.1554-1555, 2009.

[29] Yilmaz A., Javed O., and Shah M., "Object tracking: a survey," *ACM Computing Surveys*, vol.**38**, Article 13, 2006.

[30] Khan S.M., and Shah M., "A multiview approach to tracing people in crowded scenes using a planar homography constraint," *ECCV'06*, vol.**4**, pp.133-146, 2006.

[31] Khan S.M., and Shah M., "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.**31**, pp.505-519, 2009.

[32] Munkres J., "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol.**5**, pp.32-38, 1957.

[33] Devaux F.-O., Delannay D., and De Vleeschouwer C., "Autonomous production of images based on distributed and intelligent sensing," *In the Event detection algorithms public deliverable of the FP7 APIDIS project*, http://www.apidis.org/publications.htm, 2010.

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

[35] Homepage of APIDIS project: http://www.apidis.org/
Supplemental Data:
http://www.jaist.ac.jp/~chen-fan/apidis/www/results-tmm.htm
Subjective Evaluation Page:
http://www.jaist.ac.jp/~chen-fan/apidis/www/subjective-test-tmm.html
[36] De Vleeschouwer C., Chen F., Delannay D., Parisot C., Chaudy C., Martrou E., and Cavallaro A., "Distributed video acquisition and annotation for sport-event summarization," *NEM summit'08*, 2008.

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.

Damien Delannay received the M.S. and Ph.D. degree in electrical engineering from the Université catholique de Louvain (UCL), Belgium in 1998 and 2004 respectively. His first research activities dealt with digital watermarking of multimedia contents. Between 2004 and 2008, he has designed and developed stereo-vision systems for an industrial partner (ACIC), in the context of video-surveillance application such as people counting/detection. His current research interests are related to object detection and tracking in multiview scene analysis.

Christophe De Vleeschouwer is a permanent Research Associate of the Belgian NSF and an Assistant 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. He contributed to MPEG bodies, and several European projects. He now coordinates the FP7-216023 APIDIS European project (www.apidis.org), and the WALCOMO Walloon region project, respectively dedicated to video analysis for autonomous content production, and to personalized and interactive mobile video streaming.