

Multi-sensored vision for autonomous production of personalized video summaries

Fan Chen, D. Delannay, and C. De Vleeschouwer

ICTEAM, Université catholique de Louvain, Belgium
{damien.delannay, christophe.devleeschouwer}@uclouvain.be

Abstract. Democratic and personalized production of multimedia content is a challenge for content providers. In this paper, members of the FP7 APIDIS consortium explain how it is possible to address this challenge by building on computer vision tools to automate the collection and distribution of audiovisual content. In a typical application scenario, a network of cameras covers the scene of interest, and distributed analysis and interpretation of the scene are exploited to decide what to show or not to show about the event, so as to edit a video from a valuable subset of the streams provided by each individual camera. Generation of personalized summaries through automatic organization of stories is also considered. In final, 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 for small-audience, or automatic log in of annotations.

Keywords: Automatic production, personalized summarization, multi-camera.

1 Introduction

This 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. To address such kind of demands, this paper presents a unified framework for cost-effective and autonomous generation of sport team video contents from multi-sensored data. 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 will explain how that knowledge can be exploited to automate the production and personalize the summarization of video contents.

In more details, 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, 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, our proposed framework introduces three fundamental concepts, i.e. "completeness", "smoothness"

and “fineness”, to abstract the semantic and narrative requirement of video contents. We formulate the selection of temporal segments and corresponding viewpoints in the edited summary as two independent optimization problems that aim at maximizing individual user preferences (e.g. in terms of preferred player or video access resolution), given the outcomes of scene analysis algorithms. We refer to the research outputs of the FP7 APIDIS research project to demonstrate our framework.

2 Player tracking and sport action understanding

This section explains how multi-view analysis can support team sport actions monitoring and understanding. 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.

2.1 Multi-view Player 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 Khan and Shah [6] or Fleuret et al. [5]. Since all players have similar appearance in a team sport context, we focus on 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. [4].

Similar to [5],[6], our approach computes foreground likelihood independently on each view, using standard background modeling techniques. It then fusions those likelihoods by projecting them on the ground plane, thereby defining a set of so-called ground occupancy masks (see Fig.1). 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 [4] for more details. Once players and referee have been localized, histogram analysis is performed to assign a team label to each detection (see bounding boxes color in Fig. 1). 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 [4].

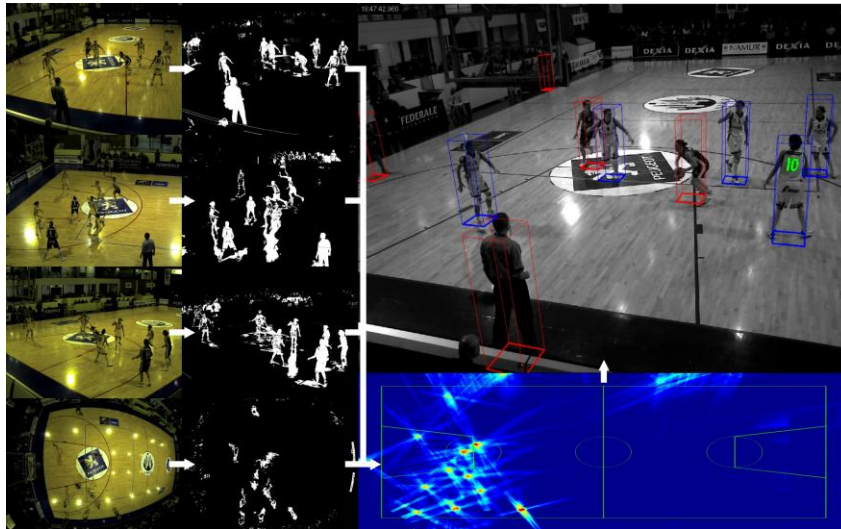


Fig. 1 – Players detection&recognition: On the left, the foreground likelihoods are extracted in each view. They are projected to define a ground occupancy map, from which people positions are extracted through an occlusion-aware greedy process.

Since the player digit can only be read when the player's back faces one of the cameras, we have to track the detected players across time. Therefore, we propagate tracks over a 1-frame horizon, based on the Munkres general assignment algorithm [7]. 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.

2.2 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, 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. Hence, we assume an accurate monitoring of the 24" clock, and of the scoreboard, and propose to organize the actions hierarchically, as a function of the observed clock and scoreboard status. This results in the tree structure depicted in Fig. 2 and 3. 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. 2 - 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. [9]. The approach achieves above 90 % accuracy.

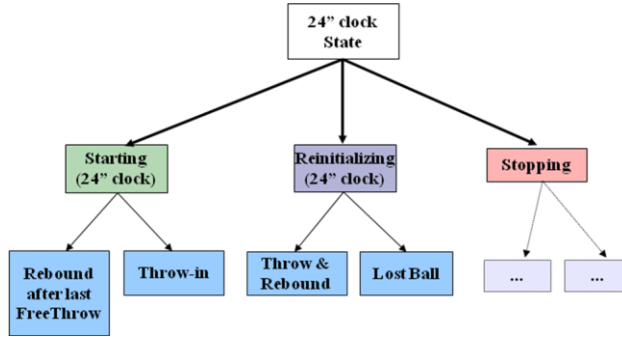


Fig. 2 – Basket-ball action tree structure.

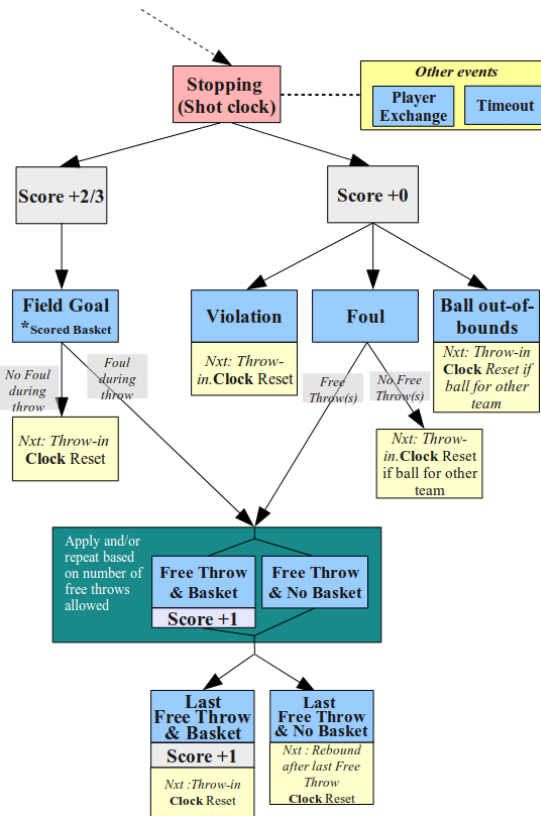


Fig. 3 – Basket-ball actions tree structure.

3 Autonomous production of personalized video summaries

To produce condensed video reports of a sport event, the temporal segments corresponding to actions that are worth being included in the summary have to be selected. For each segment, local story organization and selection of appropriate viewpoints to render the scene are also essential. In an autonomous system, all those steps have to be run in an integrated manner, independently of any human intervention. This section describes how to design and integrate video analysis, production, and summarization technologies to automate and personalize the generation of video summaries in a team sport environment, using a distributed set of cameras.

3.1 Problem and solution overview

Although the perception of a production strategy is subjective and relative to individual's perspective, there is a set of general principles whose implementation results in improved and more enjoyable viewing experience. In our proposed framework, we identify three factors affecting the quality of the produced content, and interpret production and summarization as optimization processes that trade-off among these three factors. In more details, the factors are defined as follows:

- **Completeness** 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 complete summary includes more key actions.
- **Smoothness** 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, see Owens [8].
- **Fineness** 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.

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. Our methods described in Chen and De Vleeschouwer [1-3] first define quantitative metrics to reflect completeness, fineness, and closeness. They then formulate constrained optimization problems to balance those concepts. Interestingly, it appears that 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.

In addition, for improved computational efficiency, both production and summarization are envisioned in the divide and conquer paradigm. This especially makes sense since video contents intrinsically have a hierarchical structure, starting from each frame, shots (set of consecutive frames created from similar viewpoints), to semantic segments (consecutive shots logically related to the same action), and ending with the overall sequence.

Figure 4 summarizes the framework resulting from the above considerations. The event timeframe is first cut into semantically meaningful temporal segments, such as an offense/defense round of team sports. For each segment, several narrative options are considered. Each option defines a local story, which consists of multiple shots with different camera coverage. Benefits and costs are then assigned to each local story. The cost simply corresponds to the duration of the story. The benefit reflects user satisfaction (under some individual preferences¹), and measures how some general requirements, e.g., the continuity and completeness of the story, are fulfilled. Those pairs of benefits and costs are then fed into the summarization engine, which solves a conventional resource allocation problem to find the organization of local stories that achieves the highest benefit under the constrained summary length.

Interestingly, a local story not only includes shots to render the global action at hand, but also shots for explanative and decorative purposes, e.g., replays and close-up views. For some of our previous work [3] that consider the summarization of content released by the production room, shots are simply defined based on shot boundary detection algorithms, while segments results from view type sequence monitoring. Alternatively, our proposed framework also supports the autonomous selection of viewpoints to render the action, based on a set of cameras covering the scene. In that particular case, segments and shots are defined based on scene interpretation (i.e. action recognition), and the viewpoint sequence associated to each shot is computed automatically, taking into account the nature of the shots (close-up view, replay, etc) composing the narrative option, and the position of objects-of-interest, as defined by video analysis modules.

In the sequel, our framework for automatic selection of viewpoints is presented. This autonomous production framework directly relies on the knowledge of players' positions. Due to space limitation, we omit the description of the summarization resource allocation framework, but refer interested readers to our paper [3] for a detailed description.

¹ This might involve video analysis, to measure the consistency between users preferences, and actual content of the scene.

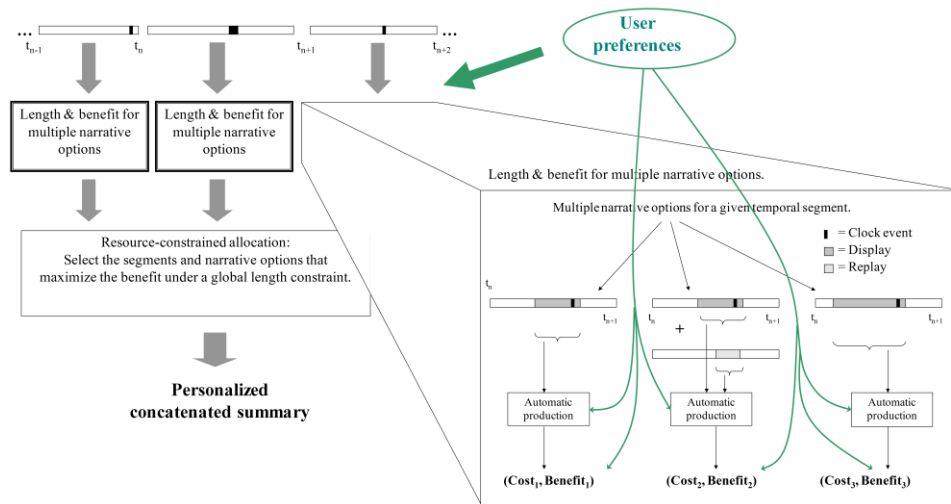


Fig. 4 - Automatic Production in Divide-and-conquer Paradigm

3.2 Viewpoint selection for team sport videos

In this section, we review our method for automatic production of video content, to render an action involving one or several players and/or objects of interest. Whilst extendable to other contexts (e.g. PTZ camera control), the process has been designed to select which fraction of which camera view should be cropped in a distributed set of still cameras to render the scene at hand in a semantically meaningful and visually pleasant way. Formal description and extensive validation of the method are provided in [1],[2]. Here, we present an intuitive description of the approach depicted in Fig. 5. Given the positions of the objects of interest, the process proceeds in three steps.

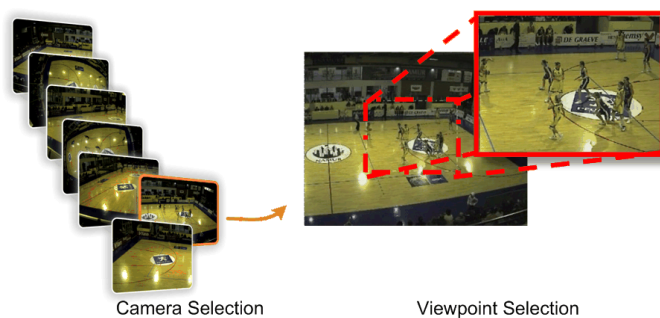


Fig. 5 - Camera selection and field of view selection.

Step 1: Camera-wise field of view selection. At each time instant and in each view, this stage selects the cropping parameters that optimize the trade-off between completeness and fineness. Here, the completeness counts the number of players in the displayed image, while the fineness measures the amount of pixels available to describe each object of interest, i.e. each player. The purpose is here to select the field-of-view that renders the scene of interest in a way that (allows the viewer to) follow the action carried out by the multiple and interacting players that have been detected, e.g. by video analysis tools.

Step 2: Frame-wise camera selection. The second stage considers the selection of the right camera to render the action at a given time instant. It rates the viewpoint selected in each view according to the quality of its completeness/closeness trade-off, and to its degree of occlusions. The highest rate corresponds to a view that (i) makes most object of interest visible, and (ii) is close to the action, meaning that it presents important objects with lots of details, i.e. a high resolution.

Step 3: Smoothing of viewpoint sequences. For the temporal segment at hand, this stage computes the parameters of an optimal virtual camera that pans, zooms and switches across views to preserve high ratings of selected viewpoints while minimizing the amount of virtual camera movements. The purpose is to build the edited video by selecting and concatenating video segments provided by multiple cameras, in a way that promotes the informative cameras, while avoiding perceptually inopportune switching between cameras and/or abrupt viewpoint changes. More details about the smoothing process are available in [1],[2].

4 Experimental results

Space limitation prevents us to include experimental results in the paper. We refer the reader to the extensive quantitative and subjective analysis published in our recent papers [1-4], but also to the demonstrations published on the web [10]. During the conference, we plan to demonstrate our real-time implementation of the integrated prototype for personalized production and summarization of pre-recorded and pre-analyzed basket-ball games.

Interestingly, the subjective experiments run based on the production component of this prototype [2] demonstrate that the viewpoints selected by the automatic virtual director is regularly preferred to the ones selected by a human producer. This is partly explained by the severe load imposed to the human operator when the number of camera increases². Hence, beyond the personalization capabilities authorized by the possibility to repeat the automatic process with different parameters, the present framework also alleviates the bottleneck experienced by a human operator, when jointly and simultaneously processing a large number of source cameras.

² In conventional systems, the load is split between several cameramen (one per camera), and one or several producers (each one selecting the best view among a subset of the cameras).

5 Conclusions

The framework presented in this paper for producing personalized video summaries has been designed to offer 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.) Story-telling complying with production principles. On the one hand, production cares about smooth camera movement while capturing the essence of team 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 action 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.

References

1. Chen F., and De Vleeschouwer C., 2009. Autonomous production of basket-ball videos from multi-sensored data with personalized viewpoints, The 10th international workshop for multimedia interactive services, pp.81-84, London, UK.
2. Chen F., and De Vleeschouwer C., 2009. Personalized production of team sport videos from multi-sensored data under limited display resolution, Computer Vision and Image Understanding, Special Issue on Sensor Fusion, 114(6), 667-680, 2010.
3. Chen F., and De Vleeschouwer C., 2009. A resource allocation framework for summarizing team sport videos, IEEE International Conference on Image Processing, pp.4349-4352, Cairo, Egypt.
4. Delannay D., Danhier N., and De Vleeschouwer C., 2009. Detection and recognition of sports (wo)men from multiple views, 3rd ACM/IEEE International Conference on Distributed Smart Cameras, Como, Italy.
5. Fleuret F., Berclaz J., Lengagne R., and Fua P., 2008. Multi-camera people tracking with a probabilistic occupancy map, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), pp. 267–282.
6. Khan S.M., and Shah M., 2009. Tracking multiple occluding people by localizing on multiple scene planes, IEEE Trans. on Pattern Analysis and Mach. Intel., 31(3), 505-519.
7. Munkres J., 1957, “Algorithms for the assignment and transportation problems,” in SIAM J. Control, vol. 5, pp. 32–38.
8. Owens J., 2007. Television sports production, 4th Ed., Burlington,MA, USA: Focal Press.
9. F.-O. Devaux, D. Delannay, and C. De Vleeschouwer, 2010. 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>.
10. APIDIS website, www.apidis.org, including some preliminary results presented during IBC2009 <http://thetis.tele.ucl.ac.be/Apidis/chen/www/results-ibc2009.html> .

Acknowledgments The author would like to thank the APIDIS partners for their support in the acquisition of the video material exploited in this work. They also thank the European Commission and the Walloon Region for funding part of this work through the FP7 APIDIS and WIST2 WALCOMO projects, respectively.