

# Multi-sensored Vision for Autonomous Production of Personalized Video Summary

Fan Chen, Damien Delannay, Christophe De Vleeschouwer, and Pascaline Parisot

*Communications and remote sensing laboratory, Université Catholique de Louvain, Belgium.*

## ABSTRACT

This chapter provides a survey of the major research efforts that have exploited computer vision tools to extend the content production industry towards automated infrastructures allowing contents to be produced, stored, and accessed at low cost and in a personalized and dedicated way.

## 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.

Hence, democratic and personalized production of multimedia content is one of the most exciting challenges that content providers will have to face in the near future. In this chapter, we explain how it is possible to address this challenge by building on computer vision tools to automate the collection and distribution of audiovisual contents.

In a typical application scenario, as depicted in Figure 1, the sensor network for media acquisition is composed of (microphones and) 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 from the streams provided by each individual camera, or interpolated from multiple cameras. The process involves numerous integrated technologies and methodologies, including but not limited to automatic scene analysis, camera viewpoint selection and control, and generation of summaries through automatic organization of stories. Considering the problem in a multi-camera environment not only mitigates the difficulty of scene understanding caused by reflection, occlusion and shadow in the single view case, but also offers higher flexibility in producing visually pleasant video reports. In final, multi-camera autonomous production/summarization can provide practical solutions to a wide range of applications, such as personalized access to local sport events through a web portal or a mobile hand-set (APIDIS, 2008; Papaoulakis, 2008), cost-effective and fully automated production of content dedicated to small-audience, e.g. souvenirs DVD, university lectures, conference (Rui, 2001; Al-Hames, 2007), etc, and interactive browsing and automated summarization for video surveillance (Yamasaki, 2008).

From a technical perspective, this chapter will present a unified framework for cost-effective and autonomous generation of video contents from multi-sensored data. It will first investigate 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.

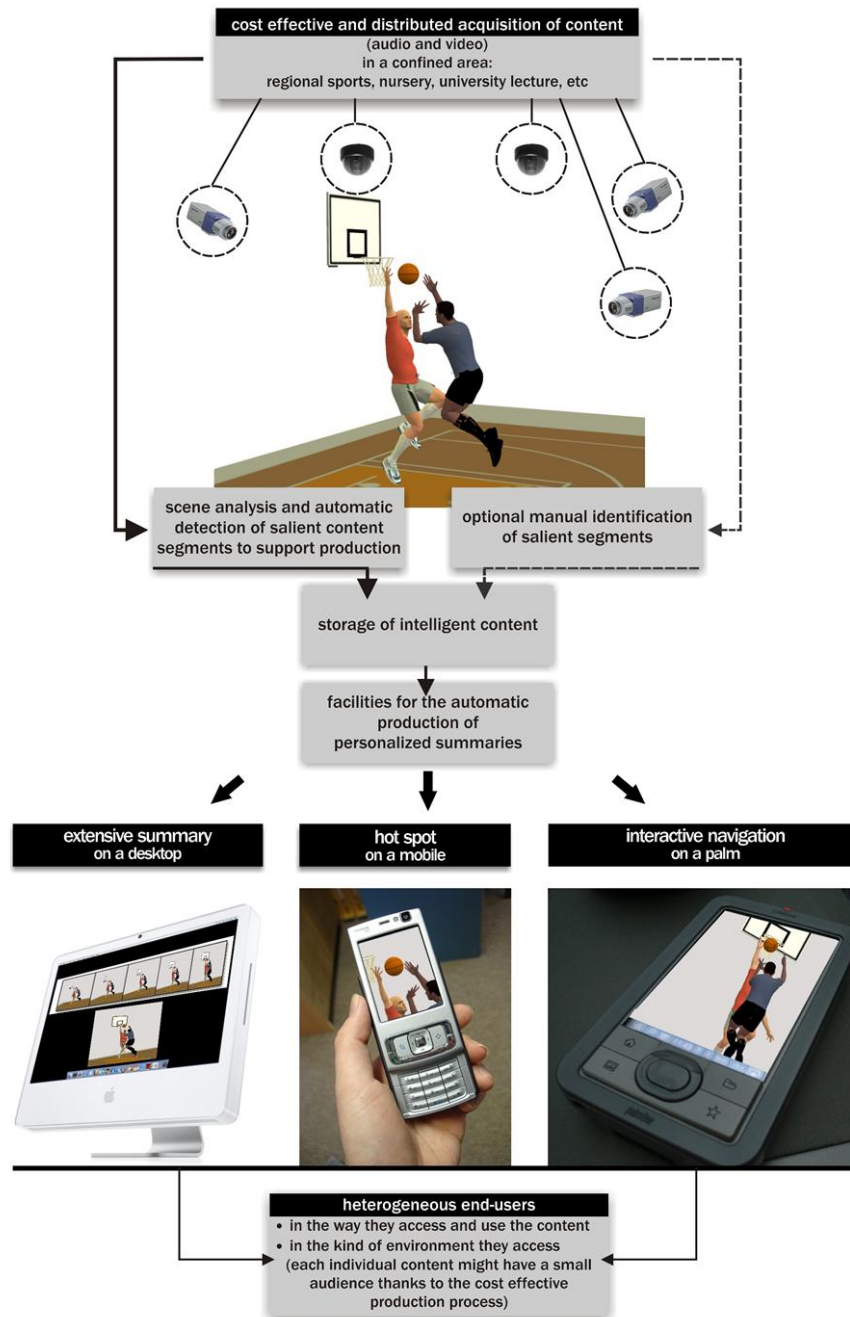


Figure 1. Vision of Autonomous Production of Personalized Video Summaries.

In more details, to identify salient segments in the raw video content, multi-camera analysis is considered, with an emphasis on people detection methods relying on the fusion of the foreground likelihood information computed in each view. We will observe that multi-view analysis can overcome traditional hurdles such as occlusions, shadows and changing illumination. This is in contrast with single sensor signal analysis, which is often subject to interpretation ambiguities, due to the lack of accurate model of the scene, and to coincidental adverse scene configurations (Delannay, 2009).

To produce semantically meaningful and perceptually comfortable video summaries based on the extraction or interpolation of images from the raw content, our proposed framework introduces three fundamental concepts, i.e. “completeness”, “smoothness” and “fineness”, to abstract the semantic and narrative requirements of video contents. Based on those concepts, as a key contribution, we formulate the selection of camera viewpoints and that of temporal segments in the summary as two independent optimization problems. In short, those problems define and trade-off the above concepts as a function of the computer vision analysis outcomes, in a way that is easily parameterized by individual user preferences. Interestingly, the solution to the viewpoint selection problem is augmented by Markov regularization mechanisms (Chen, 2009-1; Chen, 2009-2), while the formulation of the summarization problem builds on a generic resource allocation framework (Chen, 2009-3).

To demonstrate our framework, we consider both basket-ball and soccer use cases, and rely on some of the latest research outputs of the FP7 APIDIS research project (APIDIS, 2008).

## **BACKGROUND**

In this section, we survey the main achievements related to distributed video analysis, and to autonomous production of personalized video summaries. In the meantime, we position our contributions with respect to previous works in those fields, to highlight the originality of the approaches presented in subsequent sections.

### **Related Works in Autonomous Distributed Video Analysis**

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, 2006) or (Khan, 2009). In the context of team sport event monitoring, all players have similar appearance. For this reason, in this chapter, 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.

Detection of people from the foreground likelihood information, i.e. the probability that a pixel in an image belongs to the foreground, computed in multiple views has been investigated in details in the past few years. We differentiate two classes of approaches.

On the one hand, the authors in (Khan, 2006), (Lanza, 2007), (Khan, 2009), and (Delannay, 2009) adopt a bottom-up approach, and project the points of the foreground likelihood (background subtracted silhouettes) of each view to define a ground plane occupancy mask. Specifically, the change probability maps computed in each view are warped to (a set of planes that are parallel to) the ground plane based on homographies that have been computed off-line, e.g. based on reference points calibration. The projected maps are then merged to define the patches of the ground plane for which the appearance has changed compared to the background model and according to the single-view change detection algorithm.

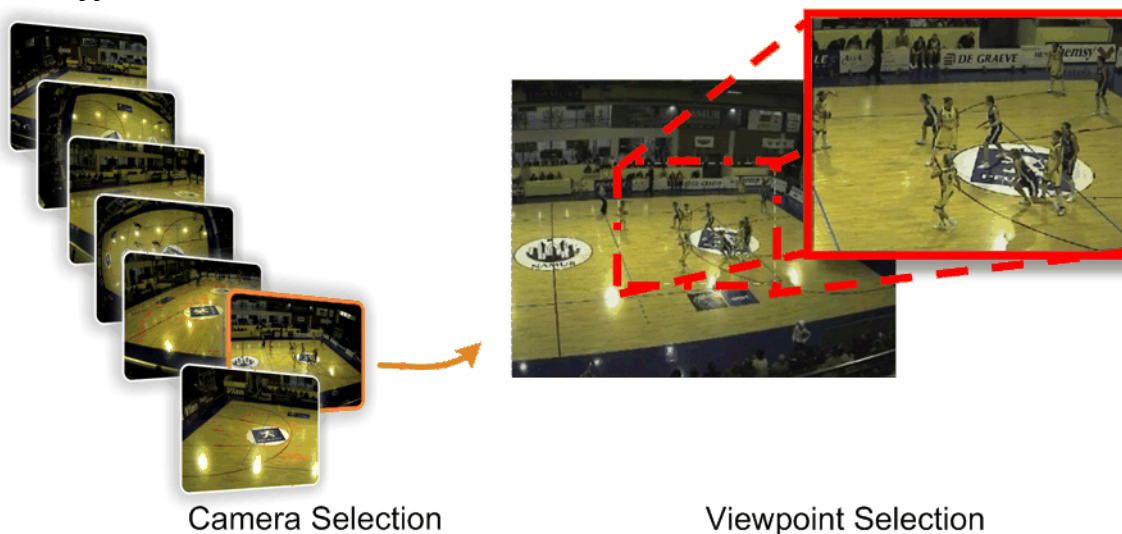
On the other hand, the works in (Berclaz, 2008), (Fleuret, 2008), and (Alahi, 2009) adopt a top-down approach. They consider a grid of points on the ground plane, and estimate the probabilities of occupancy of each point in the grid based on the back-projection of some kind of generative model in each one of the calibrated multiple views. Hence, they all start from the ground plane, and validate occupancy hypothesis based on associated appearance model in each one of the views. The approaches proposed in this second

category mainly differ based on the kind of generative model they consider (rectangle or learned dictionary), and on the way they decide about occupancy in each point of the grid (combination of multiple view-based classifiers in (Berclaz, 2008), probabilistic occupancy grid inferred from background subtraction masks in (Fleuret, 2008), and sparsely constrained binary occupancy map for (Alahi, 2009)).

The first category of methods has the advantage to be computationally efficient, since the decision about ground plane occupancy is directly taken from the observation of the projection(s) of the change detection masks of the different views. In contrast, the complexity of the second category of algorithms depends on the number of ground plane points to be investigated (chosen to limit the area to be monitored), and on the computational load associated to the validation of each occupancy hypothesis. This validation process generally involves back-projection of a 3D-world template in each one of the views. Hence, in most practical cases, the first kind of approach is significantly less complex than the second one.

Moreover, in the particular case for which the objects to detect are vertical, the methods from the first category can also exploit the entire silhouette of the object to decide about ground occupancy. This is done by projecting the foreground silhouettes on multiple parallel planes instead of on the ground plane only (Delannay, 2009), (Khan, 2009). Thereby, methods from the first category become able to achieve similar performances to the ones of the second category.

Later in this chapter, we present a player detection method that brings two fundamental improvements to methods from the first category. 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. Until now, this phenomenon had only been taken into account by the method from the second category described in (Fleuret, 2008), through a complex iterative approximation of the joint posterior probabilities of occupancy. In contrast, whilst approximate, our approach appears to be both efficient and effective.



*Figure 2. Two Key Tasks in Automatic Video Editing: Camera Selection and Viewpoint Selection*

## Related Works in Autonomous Production

Regarding the camerawork planning, we interpret the planning of “virtual” camera actions as selecting a camera view and its in-frame viewpoint, as depicted in Figure 2, 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. (Kubicek, 2005) and (Rui, 2001) 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. (Kubicek, 2005) used decaying curves to avoid fast camera switching and suppress too long shots in multimodal meetings. (Vronay, 2006-1; Vronay, 2006-2) 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. (Papaoulakis, 2008) 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 (Suh, 2003; Xie, 2006), based on some visual attention model (Itti, 1998). In contrast, we presented an automatic video production system in (Chen, 2009-1), 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.

## Related works in Personalized Video Summaries

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 (Tseng, 2003; Ferman, 2003), or solve constrained optimization of objective functions (Li, 2005; Pahalawatta, 2005). 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 (Qian, 2004; Ekin, 2004; Murphy, 2005; Jung, 2006; Pan, 2004; Gong, 2004). 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 (Gong, 2004), without taking care of story-telling artifacts. In contrast, (Albanese, 2006) considers the continuity of the clips included in the generated summary to improve story-telling, and (Chen B.W., 2009) 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 (Chen, 2009-3) 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.

## 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 production of associated content 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 the first attempt to integrate video analysis, production, and summarization technologies to automatically produce content, according to individual user preferences. We first present an overview of the proposed integrated automatic production and summarization framework. We then illustrate in details two of its main components, namely people detection from multiple views and automatic camerawork planning, in a team sport environment covered by a distributed set of still cameras.

### Problem and solution overview

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.

In our proposed framework, we identify three major factors affecting the quality of the produced summary, namely the "completeness", the "fineness" and the "smoothness", 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 story of high completeness consists of 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 (Owen, 2007).

- **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. It first defines quantitative metrics to reflect completeness, fineness, and closeness. It then formulates 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 by similar camerawork), to semantic segments (consecutive shots logically related to the identical action), and ending with the overall sequence.

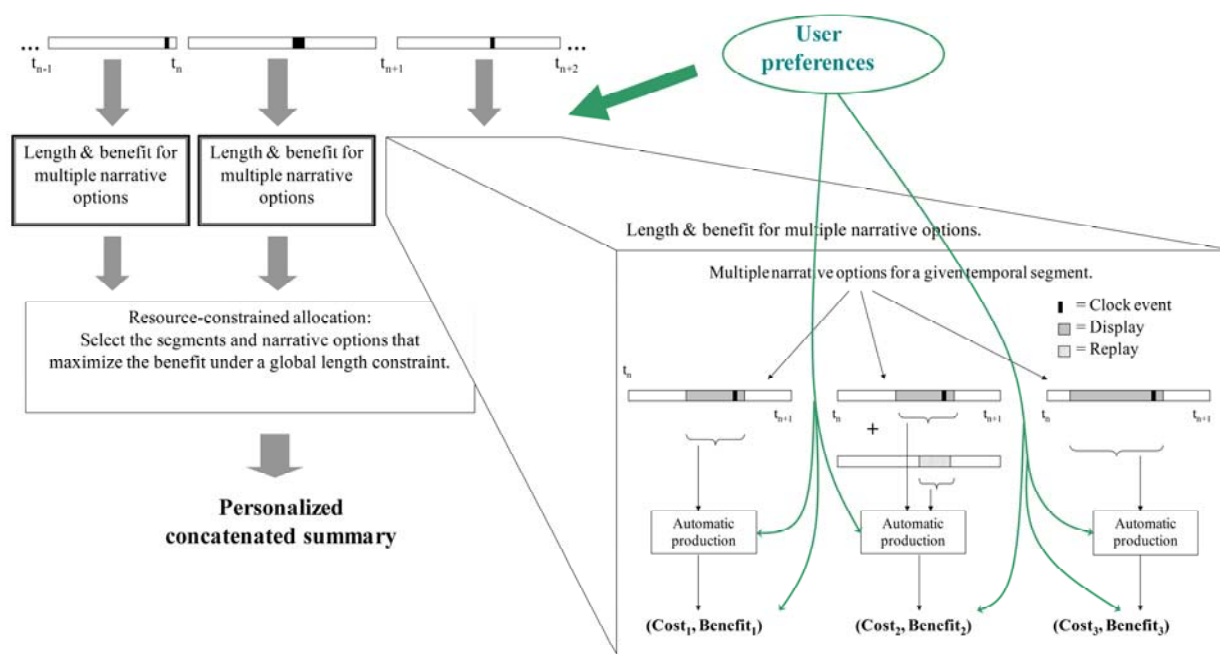


Figure 3. Automatic Production in Divide-and-conquer Paradigm

Figure 3 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, or an entry in news. For each segment, several narrative options are considered. Each option defines a local story, which consists of multiple shots with different camera coverage. 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 in sports or graph data in news. Given the timestamps and the production strategy (close-up view, replay, etc) of the shots composing a narrative option, the camerawork associated to each shot is planned automatically, taking into account the knowledge inferred about the scene by video analysis modules.

Benefits and costs are then assigned to each local story. The cost simply corresponds to the duration of the summary. The benefit reflects user satisfaction (under some individual preferences<sup>1</sup>), 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 (Everett, 1963) to find the organization of local stories that achieves the highest benefit under the constrained summary length.

In the sequel, our framework for automatic planning of camerawork is described in details and demonstrated in the context of basket-ball production. Since our production framework relies on the knowledge of players' positions, we also derive an original multi-view algorithm that detects people from their background-subtracted silhouettes.

Due to space limitation, we omit the description of the summarization resource allocation framework, but refer interested readers to our paper (Chen, 2009-3) for a detailed description and a study of the football use case.

## Camerawork Planning for Team Sport Videos

In this section, we develop an algorithm for basketball video production, as a realistic implementation of the above integrated framework for content production. 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 by assuming the knowledge of players' positions in (Chen, 2009-1; Chen 2009-2). In Figure 4, we schematically depict the three steps composing the process, and describe them as follows.

### Step 1: Camera-wise Viewpoint Selection.

At each time instant and in each view, we assume that the players' supports are known, and select the cropping parameters that optimize the trade-off between completeness and fineness.

Formally, a viewpoint  $\mathbf{v}_{ki}$  in the  $k^{\text{th}}$  camera view of the  $i^{\text{th}}$  frame is defined by the size  $S_{ki}$  and the center  $\mathbf{c}_{ki}$  of the window that is cropped in the  $k^{\text{th}}$  view for actual display. It has to be selected to include the objects of interest, and provide a fine, i.e. high resolution, description of those objects. If there are  $N$  salient objects in this frame, and the location of the  $n^{\text{th}}$  object in the  $k^{\text{th}}$  view is denoted by  $\mathbf{x}_{nki}$ , we select the optimal viewpoint  $\mathbf{v}_{ki}^*$ , by maximizing a weighted sum of object interests as follows:

$$\mathbf{v}_{ki}^* = \arg \max_{\{S_{ki}, \mathbf{c}_{ki}\}} \sum_{n=1}^N I_n \cdot \beta(S_{ki}, \mathbf{u}) \cdot \alpha \left( \frac{\|\mathbf{x}_{nki} - \mathbf{c}_{ki}\|}{S_{ki}} \right) \quad (1.1)$$

---

<sup>1</sup> Note that this might involve video analysis, to measure the consistency between the preferences of the users, and the actual content of the scene.

In the above equation:

- $I_n$  denotes the level of interest assigned to the  $n^{\text{th}}$  object detected in the scene. Note that assigning distinct weights to team sport players allows focusing on a preferred player, but also implies recognition of each player. Player digit recognition is for example considered in (Delannay, 2009). In the rest of the chapter, we assign a unit weight to all players, thereby producing a video that renders the global team sport action.
- The vector  $\mathbf{u}$  reflects the user constraints and preferences in terms of viewpoint resolution and camera view,  $\mathbf{u}=[u^{close} u^{res} \{u_k\}]$ . In particular, its component  $u^{res}$  defines the resolution of the output stream, which is generally constrained by the transmission bandwidth or end-user device resolution. Its component  $u^{close}$  is set to a value larger than 1, and increases to favor close viewpoints compared to large zoom-out views. The other components of  $\mathbf{u}$  are dealing with camera preferences, and are defined in the second step below.
- The function  $\alpha(\cdot)$  modulates the weights of the objects according to their distance to the center of the viewpoint, compared to the size of this window. Intuitively, the weight should be high and positive when the object-of-interest is located in the center of the display window, and should be negative or zero when the object lies outside the viewing area. Many instances are appropriate (Chen, 2009-1), among which the well-known Mexican Hat function.
- The function  $\beta(\cdot)$  reflects the penalty induced by the fact that the native signal captured by the  $k^{\text{th}}$  camera has to be sub-sampled once the size of the viewpoint becomes larger than the maximal resolution  $u^{res}$  allowed by the user. This function typically decreases with  $S_{ki}$ . An appropriate choice consists in setting the function equal to one when  $S_{ki} < u^{res}$ , and in making it decrease afterwards. An example of  $\beta(\cdot)$  is defined by:

$$\beta(S_{ki}, \mathbf{u}) = \left[ \min\left(\frac{u^{res}}{S_{ki}}, 1\right) \right]^{u^{close}}, \quad (1.2)$$

where  $u^{close} > 1$  increases to favor close viewpoints compared to large zoom-out views.

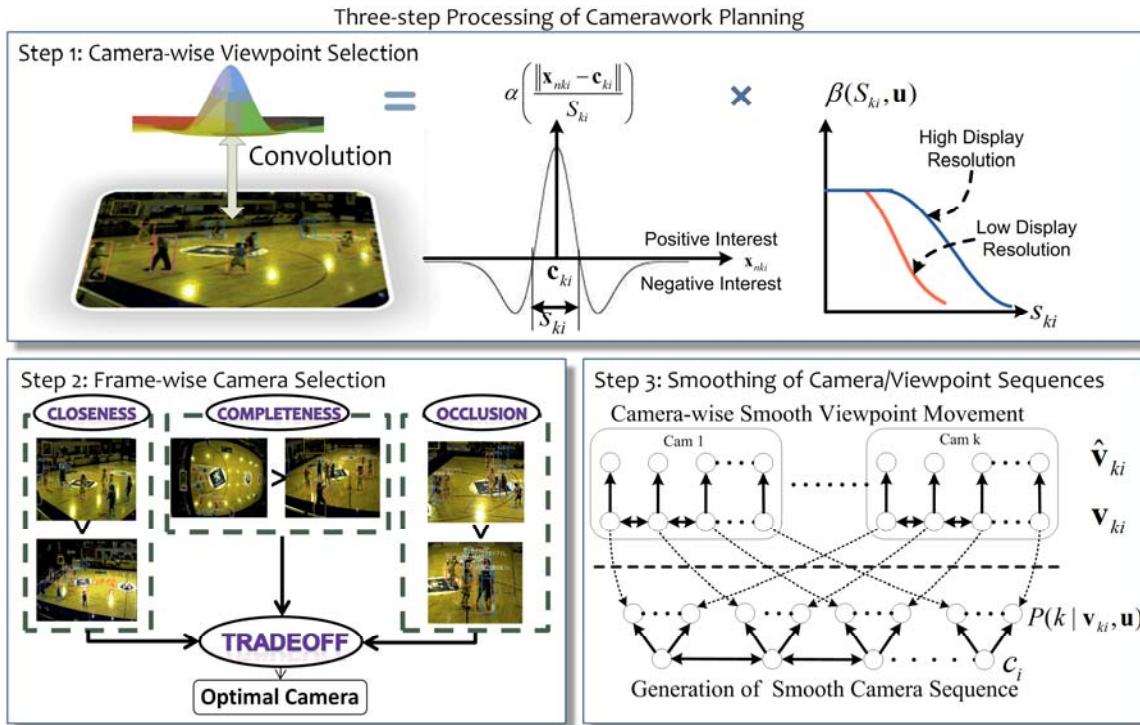


Figure 4. A Three-step Implementation of the Above Production Framework

## Step 2: Frame-wise Camera Selection

We rate 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 should correspond to a view that (1) makes most object of interest visible, and (2) is close to the action, meaning that it presents important objects with lots of details, i.e. a high resolution.

Formally, given the interest  $I_n$  of each player, the rate  $I_{ki}(\mathbf{v}_{ki}, \mathbf{u})$  associated to each camera view is defined as follows:

$$I_{ki}(\mathbf{v}_{ki}, \mathbf{u}) = u_k \cdot \sum_{n=1}^N I_n \cdot o_k(\mathbf{x}_{nki} | \bar{\mathbf{x}}) \cdot h_k(\mathbf{x}_{nki}) \cdot \beta(S_{ki}, \mathbf{u}) \cdot \alpha\left(\frac{\|\mathbf{x}_{nki} - \mathbf{c}_{ki}\|}{S_{ki}}\right) \quad (1.3)$$

In the above equation:

- $u_k$  denotes the weight assigned to the  $k^{\text{th}}$  camera, while  $\alpha$  and  $\beta$  are defined as in the first step above.
- $o_k(\mathbf{x}_{nki} | \bar{\mathbf{x}})$  measures the occlusion ratio of the  $n^{\text{th}}$  object in camera view  $k$ , knowing the position of all other objects. The occlusion ratio of an object is defined to be the fraction of pixels of the object that are hidden by other objects when projected on the camera sensor.
- The height  $h_k(\mathbf{x}_{nki})$  is defined to be the height in pixels of the projection in view  $k$  of a six feet tall vertical object located in  $\mathbf{x}_{nki}$ . Six feet is the average height of the players. The value of  $h_k(\mathbf{x}_{nki})$  is directly computed based on camera calibration. When calibration is not available, it can be estimated based on the height of the object detected in view  $k$ .

## Step 3: Smoothing of Camera/Viewpoint Sequences.

For the temporal segment at hand, we then compute 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 smoothing process is implemented based on the definition of two Markov Random Fields, as shown in Figure 4. At first, we take  $\hat{\mathbf{v}}_{ki}$  as observed data on the  $i^{\text{th}}$  image, and assume that they are noise-distorted outputs of some underlying smooth results  $\mathbf{v}_{ki}$ . Given the smooth viewpoint sequence recovered for each camera, we then compute camera-gains  $I_{ki}(\mathbf{v}_{ki}, \mathbf{u})$  of those derived viewpoints, and infer a smooth camera sequence from the second Markov field, by making the probabilities  $P(k|\mathbf{v}_{ki}, \mathbf{u})$  of each camera proportional to the gains  $I_{ki}(\mathbf{v}_{ki}, \mathbf{u})$ .

More details about the smoothing process are available in (Chen, 2009-1).

Compared to simple Gaussian smoothing filters, the depicted model enables adaptive smoothing by setting different smoothing strength on each individual frame. Furthermore, iterative slight smoothing in our method is able to achieve softer results than one-pass strong smoothing.

## Multi-view Player Detection and Recognition

As explained above, autonomous production of visual content relies on the detection (and recognition) of object-of-interest in the scene. In this section, we explain how players can be detected based on joint processing of multiple views.

The method is depicted in Figure 5. Similar to (Khan, 2009) or (Fleuret, 2008), our approach computes foreground likelihood independently on each view, using standard background modeling techniques. Our method then fusions 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 (Khan, 2009) comes both from the efficient computation of the ground occupancy mask associated to each view, and from the way

those masks are combined and processed to infer the actual position of players. In final, our method appears to improve the state of the art both in terms of computational efficiency and detection reliability.

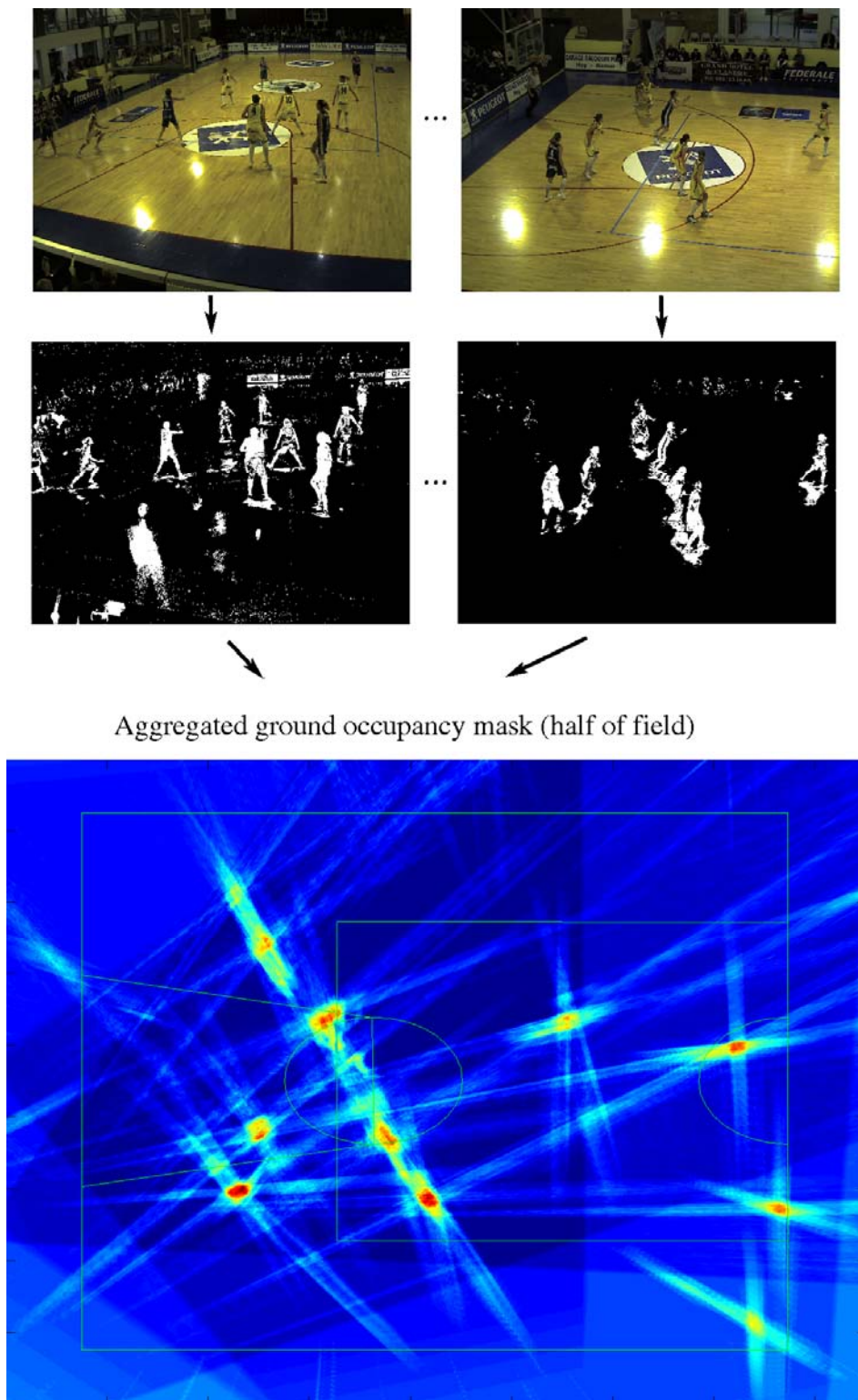


Figure 5. Multi-view People Detection. Foreground masks are projected and aggregated to define a ground plane occupancy map, from which players' positions are directly inferred.

Formally, the computation of the ground occupancy mask  $\mathbf{G}_k$  associated to the  $k^{\text{th}}$  view is described as follows. At a given time, the  $k^{\text{th}}$  view is the source of a foreground likelihood image  $\mathbf{F}_k \in [0,1]^{M_k}$ , where  $M_k$  is the number of pixels of camera  $k$ ,  $0 < k < C$ . Due to the player verticality assumption, vertical line segments anchored in occupied positions on the ground plane support a part of the detected object, and thus back-project on foreground silhouettes in each camera view. Hence, to reflect ground occupancy in  $\mathbf{x}$ , the value of  $\mathbf{G}_k$  in  $\mathbf{x}$  is defined to be the integration of the (forward-)projection of  $\mathbf{F}_k$  on a vertical segment anchored in  $\mathbf{x}$ . Obviously, this integration can equivalently be computed in  $\mathbf{F}_k$ , along the back-projection of the vertical segment anchored in  $\mathbf{x}$ . This is in contrast with (Khan, 2009), which computes the mask by aggregating the projections of the foreground likelihood on a set of planes that are parallel to the ground.

To speed up the computations associated to our formulation, we observe that, through appropriate transformation of  $\mathbf{F}_k$ , it is possible to shape the back-projected integration domain so that it also corresponds to a vertical segment in the transformed view, thereby making the computation of integrals particularly efficient through the principle of integral images. Figure 6 illustrates that specific transformation for one particular view. The transformation has been designed to address a double objective. First, points of the 3D space located on the same vertical line have to be projected on the same column in the transformed view (vertical vanishing point at infinity). Second, vertical objects that stand on the ground and whose feet are projected on the same horizontal line of the transformed view have to keep same projected heights ratios. Once the first property is met, the 3D points belonging to the vertical line standing above a given point from the ground plane simply project on the column of the transformed view that stands above the projection of the 3D ground plane point. Hence,  $\mathbf{G}_k(\mathbf{x})$  is simply computed as the integral of the transformed view over this vertical back-projected segment. Preservation of height along the lines of the transformed view even further simplifies computations.

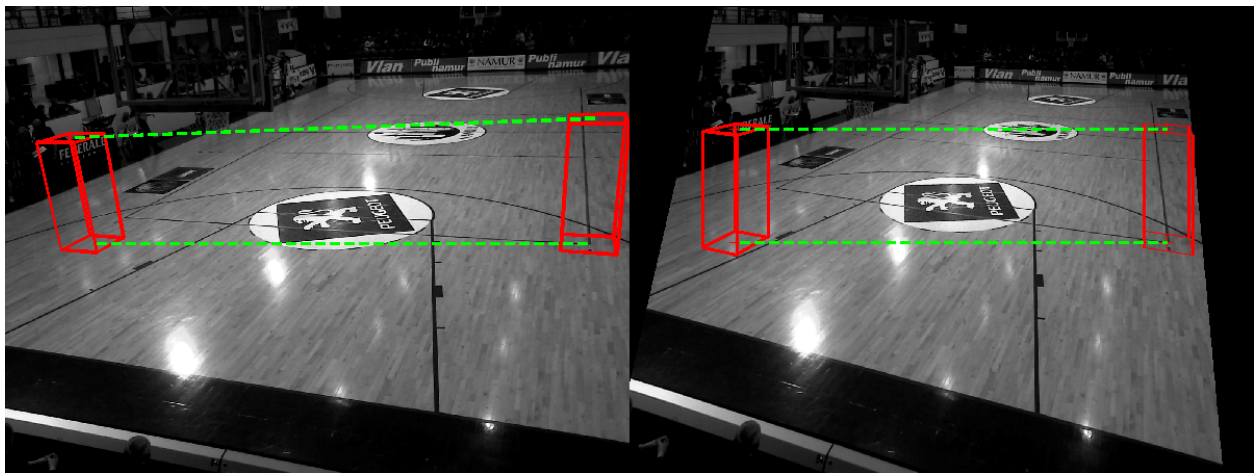


Figure 6. Efficient computation of the ground occupancy mask: the original view (on the left) is mapped to a plane through a combination of homographies that are chosen so that (1) verticality is preserved during projection from 3D scene to transformed view, and (2) ratio of heights between 3D scene and projected view is preserved for objects that lies on the same line in the transformed view.

For side views, these two properties can be achieved by virtually moving (through homography transforms) the camera viewing direction (principal axis) so as to bring the vertical vanishing point at infinity and ensure horizon line is horizontal. For top views, the principal axis is set perpendicular to the ground and a polar mapping is performed to achieve the same properties. Note that in some geometrical configurations, these transformations can induce severe skewing of the views.

Given the ground occupancy masks  $\mathbf{G}_k$  for all views, we now explain how to infer the position of the people standing on the ground. A priori, in a team sport context, we know that (i) each player induces a dense cluster on the sum of ground occupancy masks, and (ii) the number of people to detect is equal to a known value  $N$ , e.g.  $N = 12$  for basket-ball (10 players + 2 referees).

For this reason, in each ground location  $\mathbf{x}$ , we consider the sum of all projections -normalized by the number of views that actually cover  $\mathbf{x}$ -, and look for the higher intensity spots in this aggregated ground occupancy mask. To locate those spots, we have first considered a naive greedy approach that is equivalent to an iterative matching pursuit procedure. At each step, the matching pursuit process maximizes the inner product between a translated Gaussian kernel, and the aggregated ground occupancy mask. The position of the kernel which induces the larger inner-product defines the player position. Before running the next iteration, the contribution of the Gaussian kernel is subtracted from the aggregated mask to produce a residual mask. The process iterates until sufficient players have been located.

This approach is simple, but suffers from many false detections at the intersection of the projections of distinct players silhouettes from different views. This is due to the fact that occlusions induce non-linearities in the definition of the ground occupancy mask<sup>2</sup>. Hence, knowledge about the presence of some people on the ground field affects the informative value of the foreground masks in these locations. In particular, if the vertical line associated to a position  $\mathbf{x}$  is occluded by/occludes another player whose presence is very likely, this particular view should not be exploited to decide whether there is a player in  $\mathbf{x}$  or not.

For this reason, we propose to refine our naive approach as follows. To initialize the process, we define  $\mathbf{G}_k^1(\mathbf{x}) = \mathbf{G}_k(\mathbf{x})$  to be the ground occupancy mask associated to the  $k^{\text{th}}$  view, and set  $w_k^1(\mathbf{x})$  to 1 when  $\mathbf{x}$  is covered by the  $k^{\text{th}}$  view, and to 0 otherwise.

Each iteration is then run in two steps. At iteration  $n$ , the first step searches for the most likely position of the  $n^{\text{th}}$  player, knowing the position of the  $(n-1)$  players located in previous iterations. The second step updates the ground occupancy masks of all views to remove the contribution of the newly located player.

Formally, the first step of iteration  $n$  aggregates the ground occupancy mask from all views, and then searches for the denser cluster in this mask. Hence, it computes the aggregated mask as:

$$G^n(\mathbf{x}) = \frac{\sum_{k=1}^C w_k^n(\mathbf{x}) \cdot G_k^n(\mathbf{x})}{\sum_{k=1}^C w_k^n(\mathbf{x})}, \quad (1.4)$$

and then defines the most likely position  $x_n$  for the  $n^{\text{th}}$  player by

$$x_n = \underset{y}{\operatorname{argmax}} \langle G^n, \psi(y) \rangle \quad (1.5)$$

where  $\psi(y)$  denotes a Gaussian kernel centered in  $\mathbf{y}$ , and whose spatial support corresponds to the typical width of a player.

In the second step, the ground occupancy mask of each view is updated to account for the presence of the  $n^{\text{th}}$  player. In the ground position  $\mathbf{x}$ , we consider that the typical support of a player silhouette in view  $k$  is a rectangular box of width  $W$  and height  $H$ , and observe that the part of the silhouette that occludes or is occluded by the newly detected player does not bring any information about the potential presence of a player in position  $\mathbf{x}$ . In (Delannay, 2009), we estimate the fraction  $\varphi_k(\mathbf{x}, \mathbf{x}_n)$  of the silhouette in ground

---

<sup>2</sup> In other words, the ground occupancy mask of a group of players is not equal to the sum of ground occupancy masks projected by each individual player.

position  $\mathbf{x}$  that becomes non-informative in the  $k^{\text{th}}$  view, as a consequence of the presence of a player in  $\mathbf{x}_n$ . We then propose to update the ground occupancy mask and aggregation weight of the  $k^{\text{th}}$  camera in position  $\mathbf{x}$  as follows:

$$G_k^{n+1}(\mathbf{x}) = \max\left(0, G_k^n(\mathbf{x}) - \varphi_k(\mathbf{x}, \mathbf{x}_n) \cdot G_k^1(\mathbf{x}_n)\right), \quad (1.6)$$

$$w_k^{n+1}(\mathbf{x}) = \max\left(0, w_k^n(\mathbf{x}) - \varphi_k(\mathbf{x}, \mathbf{x}_n)\right). \quad (1.7)$$

For improved computational efficiency, we limit the positions  $\mathbf{x}$  investigated in the refined approach to the 30 local maxima that have been detected by the naive approach.

For completeness, we note that the above described update procedure omit the potential interference between occlusions caused by distinct players in the same view. However, the consequence of this approximation is far from being dramatic, since it ends up in omitting part of the information that was meaningful to assess the occupancy in occluded positions, without affecting the information that is actually exploited. Taking those interferences into account would require to back-project the player silhouettes in each view, thereby tending towards a computationally and memory expensive top-down approach such as the one presented in (Fleuret, 2008) and (Alahi, 2009). In these approaches, the authors propose formulations that simultaneously search for the  $N$  positions that best explain the multiple foreground masks observations. However, jointly considering all positions increases the dimensionality of the problem, and dramatically impacts the computational load. Since our experimental results show that our proposed method does not suffer from the usual weaknesses of greedy algorithms, such as a tendency to get caught in bad local minima, we believe that it compares very favorably to any joint formulation of the problem, typically solved based on iterative proximal optimization techniques. This statement is for example confirmed when comparing the results reported in (Delannay, 2009) and (Alahi, 2009).

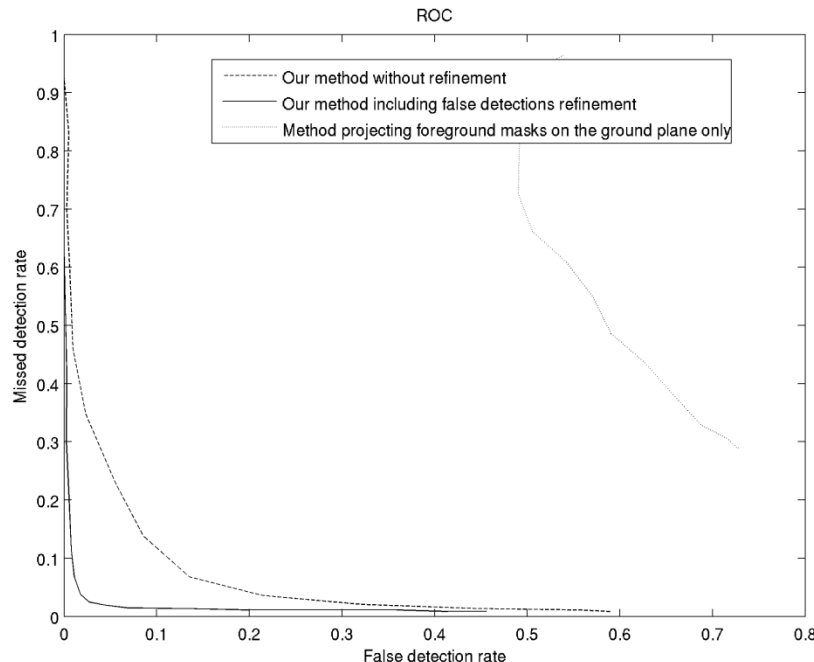


Figure 7. ROC Analysis of Player Detection Performance

## Experimental Results

In Figure 7, we plotted the average missed detection and false detection rates that are achieved by our method on a 3 minutes-long segment of the APIDIS basket ball dataset. Distinction is made between the

performance of our basic approach - which can be related to the approach proposed by (Khan, 2009) - and the improvement brought by our refined approach. We also plotted the results that we obtained when projecting the foreground masks on the ground plane only, similar to the approach described in (Khan, 2006). The achieved performance is quite satisfying with respect to the automatic production process requirements. Using these results as input, a tracking algorithm can further improve the performance assuming temporal consistency of player tracks. Combined with a number recognition (OCR) algorithm, one can track individual players from the time they enter until they exit the court. This knowledge can then be used to infer the valuable information about the ongoing events to feed the personalized summarization process.

Figure 8 gives the thumbnails of the videos produced under three different display resolutions, based on the above player detection performance. When the resolution is low, the selected viewpoint will focus on less objects or more condensed area, e.g. side view from the far end as shown in the first column. When the resolution gets higher, the selected viewpoint will include more objects and favor wide views. Readers are invited to visit the website of APIDIS project (Apidis, 2008), to view more video results and forge their own opinions. Extensive quantitative results on system behavior and subjective evaluation can be found in (Chen, 2009-1~2009-3).

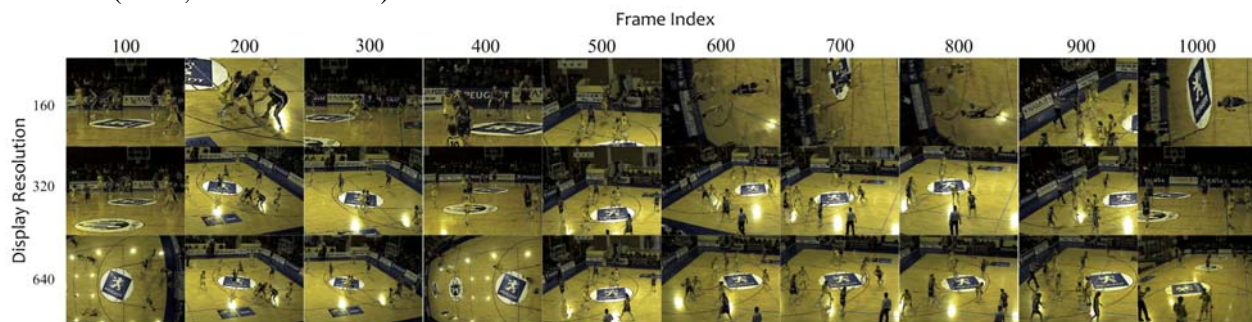


Figure 8. Results of Automatic Video Production

## FUTURE RESEARCH DIRECTIONS

The technology presented in this chapter paves the way for a novel discipline, with numerous applications ranging from coaching assistance to sport event production and summarization. Practical deployment of commercially viable systems would however benefit from advances related to:

- **Improved computational efficiency.** Real-time and low latency implementation of the players' detection algorithms would allow to control the parameters of a (set of) dynamic pan-tilt-zoom camera(s), by using the autonomous production principles to select appropriate PTZ parameters to render the scene. This would dramatically improve the quality of images compared to the ones generated based on still image cropping, which in turns would open TV broadcast markets.
- **Improved story organization.** The mechanisms controlling camerawork planning and local/global story organization are quite flexible in the way they integrate the user preferences and the applicative context. In particular, the importance assigned to a particular salient object or the benefit resulting from a local story can be arbitrarily chosen. This opens the door for a wide range of application scenarios, both within and outside sport environment.
- **Automatic collection of meta-data.** Although our frameworks of video production and summarization can live with few semantic meta-data (Chen, 2009-2; Chen, 2009-3), their personalization capabilities can be significantly refined by integrating more abundant and accurate meta-data. We expect improved automatic collection of meta-data, by making further progress on player recognition (see Delannay, 2009), ball tracking, and event recognition. Being able to generate

those metadata automatically could also open perspectives in terms of annotation of content resulting from conventional human-made production.

- **Inclusion of audio information.** Synthesis of audio commentary from the knowledge collected about the action is certainly a central task to consider in a near future. It brings benefits in terms of user experience, but also multimodal challenges related to the definition of audiovisual completeness, smoothness, and fineness.

## CONCLUSION

It appears from this chapter that 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.

## REFERENCES

Alahi A., Boursier Y., Jacques L., & Vandergheynst P., (2009). A sparsity constrained inverse problem to locate people in a network of cameras, *Proceedings of the 16th International Conference on Digital Signal Processing (DSP)*, Santorini, Greece.

Albanese M., Fayzullin M., Picariello A., & Subrahmanian V.S., (2006). The priority curve algorithm for video summarization, *Information Systems*, 31(7), 679-695.

Al-Hames M., Hornler B., Muller R., Schenk J., & Rigoll G., (2007). Automatic multi-modal meeting camera selection for video-conferences and meeting browsers, *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo*,(pp.2074-2077), Beijing, China: IEEE.

APIDIS. (2008). Autonomous Production of Images Based on Distributed and Intelligent Sensing.

Homepage of the APIDIS project.

<http://www.apidis.org/>

Demo videos related to this paper.

[http://www.apidis.org/Initial Results/APIDIS%20Initial%20Results.htm](http://www.apidis.org/Initial%20Results/APIDIS%20Initial%20Results.htm)

Berclaz J., Fleuret F., and Fua P., (2008). Principled detection-by-classification from multiple views, *Proceedings of the International Conference on Computer Vision Theory and Application (VISAPP)*, vol. 2, Funchal, Madeira, Portugal, pp. 375–382.

Chen B.W., Wang J.C., & Wang J.F., (2009). A novel video summarization based on mining the story-structure and semantic relations among concept entities, *IEEE Transactions on Multimedia*, 11(2), 295-312.

Chen F., & De Vleeschouwer C., (2009-1). Autonomous production of basket-ball videos from multi-sensed data with personalized viewpoints, *The 10th international workshop for multimedia interactive services*(pp.81-84), London, UK:IEEE.

Chen F., & De Vleeschouwer C., (2009-2). Personalized production of team sport videos from multi-sensed data under limited display resolution, *Computer Vision and Image Understanding, Special Issue on Sensor Fusion*, accepted under revision, available upon request.

Chen F., & De Vleeschouwer C., (2009-3). A resource allocation framework for summarizing team sport videos, *2009 IEEE International Conference on Image Processing*, (Accepted) Cairo, Egypt: IEEE.

Delannay D., Danhier N., & De Vleeschouwer C., (2009). Detection and recognition of sports (wo)men from multiple views, *The 3rd ACM/IEEE International Conference on Distributed Smart Cameras*, Como, Italia: IEEE.

Ekin A., & Tekalp M., (2004), Automatic soccer video analysis and summarization, US2004130567.

Everett H., (1963). Generalized lagrange multiplier method for solving problems of optimum Allocation of Resources, *Operations Research*, 11(3), 399-417.

Ferman A.M., & Tekalp A.M., (2003). Two-stage hierarchical video summary extraction to match low-level user browsing preferences, *IEEE Transactions on Multimedia*, 5(2), 244–256.

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.

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

Itti L., Koch C., & Niebur E., (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11). 1254-1259.

Jung C., Kin C., Kim S.K., Lee G., Kim W.Y., & Hwang S., (2006). Method and Apparatus for Summarizing Sports Moving Picture, SAMSUNG ELECTRONICS CO LTD, JP2006148932.

Khan S., & Shah M., (2006). A multiview approach to tracing people in crowded scenes using a planar homography constraint, *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, vol. 4, Graz, Austria, pp. 133–146.

Khan S.M., Shah M., (2009). Tracking multiple occluding people by localizing on multiple scene planes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3), 505-519.

Kubicek R., Zak P., Zemcik P., & Herout A., (2008). Automatic video editing for multimodal meetings, *International Conference on Computer Vision and Graphics 2008 (1-12)*, Warsaw, Poland: Springer.

Lanza A., Di Stefano L., Berclaz J., Fleuret F., & Fua P., (2007). Robust multiview change detection, *British Machine Vision Conference (BMVC)*, Warwick, UK.

Li Z., Schuster G.M., & Katsaggelos A.K., (2005). MINMAX optimal video summarization, *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10), 1245–1256.

Murphy N., & Smeaton A., (2005). Audio-visual sequence analysis, WO2005124686 A1, UNIV DUBLIN CITY, Publication info: IE20040412 (A1).

Owens, J., (2007), *Television sports production*, 4th Edition, Burlington, MA, USA: Focal Press.

Pahalawatta P.V., Zhu L., Zhai F., & Katsaggelos A.K., (2005). Rate-distortion optimization for internet video summarization and transmission, *IEEE 7th Workshop on Multimedia Signal Processing*. ( pp.1-4), Shanghai, China: IEEE.

Pan H., & Li B.X., (2004). Summarization of soccer video content, US20040017389A1.

Papaoulakis N., Doulamis N., Patrikakis C., Soldatos J., Pnevmatikakis A., & Protonotarios E., (2008). Real-time video analysis and personalized media streaming environments for large scale athletic events, *Proceeding of the 1st ACM Workshop on Analysis and Retrieval of Events/Actions and Workflows in Video Streams* (pp.105-112), Vancouver, Canada: The Association for Computing Machinery.

Qian R., & Haering N., (2004). Method for automatic extraction of semantically significant events from video, US6721454 (B1), SHARP LAB OF AMERICA INC.

Rui Y., Gupta A., & Cadiz J.J., (2001). Viewing meetings captured by an omni-directional camera, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp.450-457). Seattle, USA: The Association for Computing Machinery.

Suh B., Ling H., Bederson B.B., & Jacobs D.W., (2003). Automatic thumbnail cropping and its effectiveness, *Proceedings of the 16th Annual ACM Symposium on User interface Software and Technology*(pp.95-104). Vancouver, Canada:The Association for Computing Machinery.

Tseng B.L., & Smith J.R., (2003). Hierarchical video summarization based on context clustering, In J.R. Smith, S. Panchanathan, & T. Zhang (Ed.), *Internet Multimedia Management Systems IV: Proceedings of SPIE*, (pp. 14-25). Orlando, USA: SPIE-International Society for Optical Engine.

Vronay D., Wang S., Zhang D., & Zhang W., (2006-1). Automatic video editing for real-time multi-point video conferencing, *US Patent 20060251384*.

Vronay D., Wang S., Zhang D., & Zhang W., (2006-2). Automatic video editing for real-time generation of multiplayer game show videos, *US Patent 20060251383*.

Xie X., Liu H., Ma W.Y., & Zhang H.J., (2006). Browsing large pictures under limited display sizes, *IEEE Transactions on Multimedia*, 8(4), 707-715.

Yamasaki T., Nishioka Y., & Aizawa K.,(2008). Interactive retrieval for multi-camera surveillance systems featuring spatio-temporal summarization. *Proceeding of the 16th ACM international Conference on Multimedia:MM '08* (pp.797-800), New York, USA: The Association for Computing Machinery.

Yilmaz A., Javed O., Shah M., (2006). Object tracking: a survey, *ACM J. Computing Surveys*.