Computer Vision and Image Understanding 114 (2010) 667–680

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Personalized production of basketball videos from multi-sensored data under limited display resolution

Fan Chen *, Christophe De Vleeschouwer

Laboratoire de Télécommunications et Télédetection, Ecole Polytechnique de Louvain, Université Catholique de Louvain, Belgium

ARTICLE INFO

Article history:
Received 3 December 2008
Accepted 5 January 2010
Available online 22 January 2010

Keywords:
Content repurposing
Multi-sensored data fusion
Autonomous content production

ABSTRACT

Integration of information from multiple cameras is essential in television production or intelligent surveillance systems. We propose an autonomous system for personalized production of basketball videos from multi-sensored data under limited display resolution. The problem consists in selecting the right view to display among the multiple video streams captured by the investigated camera network. A view is defined by the camera index and the parameters of the image cropped within the selected camera. We propose criteria for optimal planning of viewpoint coverage and camera selection. Perceptual comfort is discussed as well as efficient integration of contextual information, which is implemented by smoothing generated viewpoint/camera sequences to alleviate flickering visual artifacts and discontinuous story-telling artifacts. We design and implement the estimation process and verify it by experiments, which shows that our method efficiently reduces those artifacts.

1. Introduction

The Autonomous Production of Images based on Distributed and Intelligent Sensing (APIDIS) project tends to provide a solution to generate personalized contents for improved visual representation of controlled scenarios such as sports television and surveillance, where image quality and perceptual comfort are as essential as efficient integration of contextual information [1]. Based on contextual information provided by conventional video analysis tools, e.g., activity detection and object tracking, we consider autonomous production of personalized video sequences from multiple fixed camera views, under limited display resolution and various user preferences. Since we consider a few cameras with quite heterogeneous lens and scene coverage, most of the state-of-the-art free-viewpoint synthesis methods produce blurred results in this situation [2,3]. We thus consider selecting a camera view and its cropping parameters, rather than synthesizing a free-viewpoint scene. Two important topics we deal with are: (i) how to select optimal viewpoints, i.e., cropping parameters in a given camera, so that they are tailored to limited display resolution and (ii) how to smooth camera/viewpoint sequences to remove production artifacts. Production artifacts consist of both visual artifacts, which mainly means flickering effects due to shaking or fast zoom in/out of viewpoints, and story-telling artifacts such as the discontinuity of story caused by fast camera switching and dramatic viewpoint movement.

Data fusion of multiple cameras has been widely discussed in the literature. Those previous works could be roughly classified into three major categories according to their various purposes. Methods in the first category deal with camera calibration and intelligent camera controlling by integrating contextual information of the multi-camera environment [4]. Reconstruction of 3D scene from multiple cameras is also a hot topic [5]. The third category uses multiple cameras to solve certain problems such as occlusion in various applications, e.g., people tracking [6] or arbitrary viewpoint video synthesis [2]. All these works focus much on the extraction of important contextual information, but consider little on visual quality of their outputs, such as those artifacts mentioned above.

Regarding autonomous video production, there are some methods proposed in the literature for selecting the most representative area from a standalone image. Suh et al. [7] defined the optimal cropping region as the minimum rectangle which contained saliency over a given threshold, where the saliency was computed by the visual attention model [8]. In Ref. [9], another attention model based method was proposed, where they discussed more the optimal shifting path of attention than the decision of viewpoint. In contrast, our method considers automatic adaptation of viewpoint size with respect to display resolution as well as determination of viewpoint center.

One most similar work is given in Ref. [10], where an automatic production system for soccer videos was proposed and viewpoint selection based on scene understanding was also discussed. How-

* Corresponding author.
E-mail addresses: fan.chen@uclouvain.be (F. Chen), christophe.devleeschouwer@uclouvain.be (C. De Vleeschouwer).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.01.005

ever, their system only switches viewpoints among three fixed shot sizes according to several fixed rules, which leads to uncomfortable visual artifacts due to dramatic changing of shot sizes. Furthermore, they only discussed the single-camera case.

Targeting the producing of semantically meaningful and perceptually comfortable contents from raw multi-sensored data, we propose a computationally efficient production system, based on the paradigm of divide-and-conquer. We summarize major factors of our target by three keywords, which are "completeness", "closeness" and "smoothness". Completeness stands for the integrity of view rendering. Closeness defines the fineness of detail description, and smoothness is a term referring to the continuity of both viewpoint movement and story telling. By trading off among those factors, we develop methods for selecting optimal viewpoints and cameras to fit the display resolution and other user preferences, and for smoothing these sequences for a continuous and graceful story-telling. There are a long list of possible user preferences, such as user's profile, user's browsing history, and device capabilities. We summarize narrative preferences into four descriptors, i.e., user preferred team, user-preferred player, user preferred event, and user preferred camera. All device constraints, such as display resolution, network speed, decoder's performance, are abstracted as the preferred display resolution and video length. Since preferences on video length and events are related to frame selection which is not considered in video production, we consider four user preferences on camera, device resolution, player and team in this paper.

The capability to take those preferences into account obviously depends on the knowledge captured about the scene through video analysis tools, e.g., detecting which team is offending or defending. However and more importantly, it is worth mentioning that our framework is generic in that it can include any kind of user preferences.

In Section 2, we explain the estimation framework of both selection and smoothing of viewpoints and camera views, and give their detailed formulation and implementation. In Section 3, more technical details are given and experiments are made to verify the efficiency of our system. We then further clarify the advantage of our system by comparing it to related works in Section 4. Finally, we conclude this work and list a number of possible paths for future research.

2. Autonomous production of personalized basketball videos from multi-sensored data

Although it is difficult to define an absolute rule to evaluate the performance of organized stories and determined viewpoints in presenting a generic scenario, production of sport videos has some general principles [11]. For basketball games, we summarize these rules into three major trade-offs.

The first trade-off arises from the personalization of the production. Specifically, it originates from the conflict between preserving general production rules of sports videos and maximizing satisfaction of user preferences. Some basic rules of video production for basketball games could not be sacrificed for better satisfaction of user preferences, e.g., the scene must always include the ball, and well balanced weighting should be taken between the dominant player and the user-preferred player when rendering an event.

The second trade-off is the balance between completeness and closeness of the rendered scene. The intrinsic interest of basketball games partially comes from the complexity of team working, whose clear description requires spatial completeness in camera coverage. However, many highlighted activities usually happen in a specific and bounded playing area. A close view emphasizing those areas increases the emotional involvement of the audience with the play, by moving the audience closer to the scene. Closeness is also required to generate a view of the game with sufficient spatial resolution under a situation with limited resources, such as small display size or limited bandwidth resources of handheld devices.

The final trade-off balances accurate pursuit of actions of interest along the time, and smoothness of viewpoint movement. The need for the audience to know the general situation regarding the game throughout the contest is a primary requirement and main purpose of viewpoint switching. When we mix angles of different cameras for highlighting or other special effects, smoothness of camera switching should be kept in mind to help the audience to rapidly re-orient the play situation after viewpoint movements [11].

Given the meta-data gathered from multi-sensor video data, we plan viewpoint coverage and camera switching by considering the above three trade-offs. We give an overview of our production framework in Section 2.1, and introduce some notations on meta-data in Section 2.2. In Section 2.3, we propose our criteria for selecting viewpoint and camera on an individual frame. Smoothing of viewpoint and camera sequences is explained in Section 2.4.

2.1. Overview of the production framework

It is unavoidable to bring discontinuity to story-telling contents when switching camera views. In order to suppress the influence of this discontinuity, we usually locate dramatic viewpoint or camera switching during the gap between two highlighted events, to avoid possible distraction of the audience from the story. Hence, we can envision our personalized production in the divide-and-conquer paradigm, as shown in Fig. 1 . The whole story is first divided into several segments. Optimal viewpoints and cameras are determined locally within each segment by trading off between benefits and costs under specified user preferences. Furthermore, estimation of optimal camera or viewpoints is performed in a hierarchical structure. The estimation phase takes bottom-up steps from all individual frames to the whole story. Starting from a standalone frame, we optimize the viewpoint in each individual camera view, determine the best camera view from multiple candidate cameras under the selected viewpoints, and finally organize the whole story. When we need to render the story to the audience, a top-down processing is taken, which first divide the video into non-overlapped segments. Corresponding frames for each segment are then picked up, and are displayed on the target device with specified cameras and viewpoints.

Intrinsic hierarchical structure of basketball games provides reasonable grounds for the above vision, and also gives clues on segment separation. As shown in Fig. 2, a game is divided by rules into a sequence of non-overlapped ball-possession periods. A ball-possession period is the period of game when the same team holds the ball and makes several trials of scoring. Within each period, several events might occur during the offence/defence process. According to whether the event is related to the 24-s shot clock, events in a basketball game could be classified as clock-events and non-clock-events. Clock-events will not overlap with each other, while non-clock-events might overlap with both clock-/non-clock-events. In general, one ball-possession period is a rather fluent period and requires the period-level continuity of viewpoint movement.

In this paper, we first define the criteria for evaluating viewpoints and cameras on each individual frame. Camera-wise smoothness of viewpoints is then applied to all frames within each ball-possession period. Based on determined viewpoints, a camera sequence is selected and smoothed.

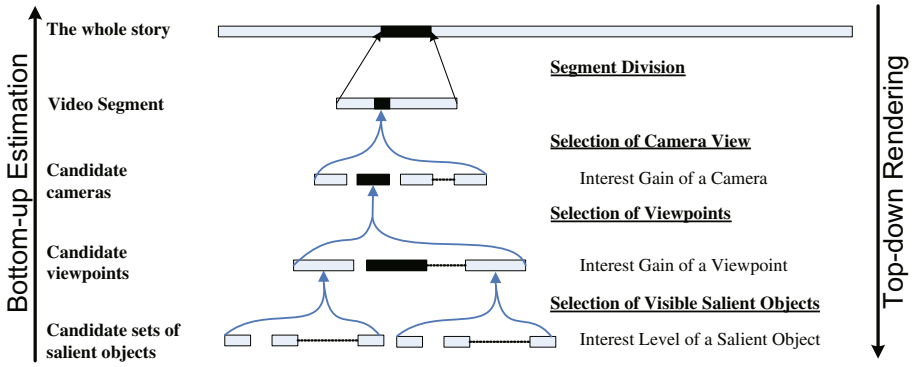


Fig. 1. Hierarchical working flow in our personalized production.

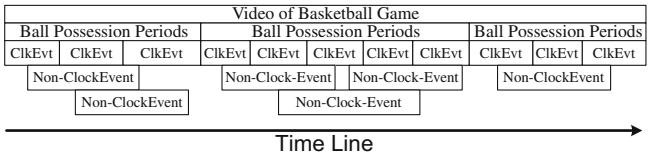


Fig. 2. Hierarchical structure of basketball videos.

2.2. Meta-data and user preference

Input data fed into our system include video data, associated meta-data, and user preferences. Let's assume that we have gathered a database of basketball video sequences, which are captured simultaneously by K different cameras. All cameras are loosely synchronized and produce the same number of frames, i.e., N frames, for each camera. On the ith frame captured at time t_i, M_i different salient objects, denoted by $\{\mathbf{o}_{im}|m=1,\ldots,M_i\}$, are detected in total from all camera views. We have two kinds of salient objects defined. The first class includes regions for players, referees, and the ball, which are used for scene understanding. The second class includes the basket, coach bench, and some landmarks of the court, which are used in both scene understanding and camera calibration. Objects of the first class are automatically extracted from the scene typically based on background subtraction algorithm [12,13], while those from the second class are manually labeled because their positions are constant on fixed cameras. In the literature, many previous works have been done on automatic detection of people [14–17]. By further utilizing the information in our controlled scenario with fixed cameras, several dedicated player detectors with high detection rate have been successfully implemented for the APIDIS project [18–20]. Results presented in this paper have been generated based on the positions detected by Ref. [18]. We define the mth salient object as $\mathbf{o}_{im}=[\mathbf{o}_{kim}|k=1,\ldots,K]$, where $\mathbf{o}_{kim}$ is the mth salient object in the kth camera.

All salient objects are represented by regions of interest. A region $\mathbf{r}$ is a set of pixel coordinates that are belonging to this region. If $\mathbf{o}_{im}$ does not appear in the kth camera view, we set $\mathbf{o}_{kim}$ to the empty set ϕ. With $\mathbf{r}_1$ and $\mathbf{r}_2$ being two arbitrary regions, we first define several elemental functions on one or two regions as

$$\textit{Area}: \quad \mathscr{A}(\mathbf{r}_1)=\sum_{\mathbf{x}\in\mathbf{r}_1}1; \tag{1}$$

$$\textit{Center}: \quad \mathscr{C}(\mathbf{r}_1)=\frac{1}{\mathscr{A}(\mathbf{r}_1)}\sum_{\mathbf{x}\in\mathbf{r}_1}\mathbf{x}; \tag{2}$$

$$\textit{Visibility}: \quad \mathscr{V}(\mathbf{r}_1|\mathbf{r}_2)=\begin{cases}1, & \mathbf{r}_1\subseteq\mathbf{r}_2;\\ -1, & \text{otherwise};\end{cases} \tag{3}$$

$$\textit{Distance}: \quad \mathscr{D}(\mathbf{r}_1,\mathbf{r}_2)=\|\mathscr{C}(\mathbf{r}_1)-\mathscr{C}(\mathbf{r}_2)\|; \tag{4}$$

which will be used in our later sections.

Furthermore, we define user preference by a parameter set $\mathbf{u}$, which includes both narrative and restrictive preferences, such as favorites and device capabilities. As mentioned in Section 1, we mainly consider four user preferences, i.e., the preferred camera, the preferred player, the preferred team and the preferred device resolution. Accordingly, $\mathbf{u}$ includes four descriptors for user preferences on camera, player, team and device resolution, respectively. The descriptors will be formally defined in Sections 2.3 and 2.4.

2.3. Selection of camera and viewpoints on individual frames

For simplicity, we put aside the smoothing problem in the first step, and start by considering the selection of a proper viewpoint on each standalone frame. We use the following two subsections to explain our solution to this problem from two aspects, i.e., evaluation of various viewpoints on the same camera view and evaluation of different camera views.

2.3.1. Computation of optimal viewpoints in each individual camera

Although evaluation of viewpoint is a highly subjective task that still lacks an objective rule, we have some basic requirements on our viewpoint selection. It should be computational efficient, and should be adaptable under different device resolutions. For a device with high display resolution, we usually prefer a complete view of the whole scene. When the resolution is limited due to device or channel constraints, we have to sacrifice part of the scene for improved representation of local details. For an object just next to the viewpoint border, it should be included to improve overall completeness of story-telling if it shows high relevance to the current event in later frames, and it should be excluded to prevent the viewpoint sequence from oscillating if it always appears around the border. In order to keep a safe area to deal with this kind of object, we prefer that visible salient objects inside the determined viewpoint are closer to the center while invisible objects should be driven away from the border of viewpoint, as far as possible.

We let the viewpoint for scene construction in the ith frame of the kth camera be $\mathbf{v}_{ki}$. Viewpoint $\mathbf{v}_{ki}$ is defined as a rectangular region. For natural representation of the scene, we limit the aspect ratio of all viewpoints to be the same aspect ratio of the display device. Therefore, for each $\mathbf{v}_{ki}$, we have only three free parameters, i.e., the horizontal center v_{kix}, the vertical center v_{kiy}, and the width v_{kiw}, to tune. Individual optimal viewpoint is obtained by maximizing the interest gain of applying viewpoint $\mathbf{v}_{ki}$ to the ith frame of the kth camera, which is defined as a weighted sum of attentional interests from all visible salient objects in that frame, i.e.,

$$\mathscr{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})=\sum_m w_{kim}(\mathbf{v}_{ki},\mathbf{u})\mathscr{I}(\mathbf{o}_{kim}|\mathbf{u}), \tag{5}$$

where $\mathscr{I}(\mathbf{o}_{kim}|\mathbf{u})$ is the interest of a salient object $\mathbf{o}_{kim}$ under user preference $\mathbf{u}$, which is mainly determined by preferred player u^{P}

and preferred team u^{TM} in $\mathbf{u}$. In the present paper, pre-defined interest function $\mathscr{I}(\mathbf{o}_{kim}|\mathbf{u})$ will give different weighting according to different values of $\mathbf{u}$, which reflects narrative user preferences. For instance, a player specified by the audience is assigned a higher interest than a player not specified, and the ball is given the highest interest so that it is always included in the scene. We will explain a practical setting of $\mathscr{I}(\mathbf{o}_{kim}|\mathbf{u})$ with more details in the next section.

We define $w_{kim}(\mathbf{v}_{ki},\mathbf{u})$ to weight the attentional significance of a single object within a viewpoint. Mathematically, we take $w_{kim}(\mathbf{v}_{ki},\mathbf{u})$ in a form as follows:

$$w_{kim}(\mathbf{v}_{ki},\mathbf{u}) = \frac{\mathscr{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})}{\ln \mathscr{A}(\mathbf{v}_{ki})} \exp\left[-\frac{\mathscr{D}(\mathbf{o}_{kim},\mathbf{v}_{ki})^2}{2[u^{\mathrm{DEV}}]^2}\right], \tag{6}$$

where we use u^{DEV} to denote limitation of current device resolution in user preference $\mathbf{u}$. Our definition of $w_{kim}(\mathbf{v}_{ki},\mathbf{u})$ consists of three major parts: the exponential part which controls the concentrating strength of salient objects around the center according to the pixel resolution of device display; the zero-crossing part $\mathscr{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ which separates positive interests from negative interests at the border of viewpoint; and the appended fraction part $\ln \mathscr{A}(\mathbf{v}_{ki})$, which calculates the density of interests to evaluate the closeness, is set as a logarithm function. Note that $\mathscr{V}(\mathbf{o}_{kim}|\mathbf{v}_{ki})$ is positive only when salient object $\mathbf{o}_{kim}$ is fully contained inside viewpoint $\mathbf{v}_{ki}$, which shows the tendency of keeping a salient object intact in viewpoint selection. As shown in Fig. 3, the basic idea of our definition is to change the relative importance of completeness and closeness by tuning the sharpness of central peak and modifying the length of tails. When u^{DEV} is small, the exponential part decays quite fast, which tends to emphasize objects closer to the center and ignore objects outside the viewpoint. When u^{DEV} gets larger, penalties for invisible objects are increased, which is on incentive to be complete and to display all salient objects. Therefore, $\mathscr{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ describes the trade-off between completeness (displaying as much objects as possible) and fineness (rendering the objects with a higher resolution) of scene description in individual frames.

A viewpoint that maximizes $\mathscr{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u})$ drives visible objects closer to the center and leads to greater separations of invisible objects from the center. We compute optimal viewpoint $\hat{\mathbf{v}}_{ki}$ individually for each frame by

$$\hat{\mathbf{v}}_{ki} = \arg\max_{\mathbf{v}_{ki}} \quad \mathscr{I}_{ki}(\mathbf{v}_{ki}|\mathbf{u}). \tag{7}$$

We perform a grid search on each frame, by first fixing the viewpoint center and then searching for its corresponding optimal viewpoint width. It is easy to prove that the optimal maximum could only occur right after including a new salient object into the viewpoint. If there are M salient objects, we only need to test M different widths to find the optimal viewpoint size for a given position. Therefore, if each image is of size $W \times H$ pixels, the overall computation complexity for processing N images is $O(NWHM)$. In our implementation, we further use a sparse grid with cell size being 16×12 pixels and restrict the searching area of viewpoint center to the minimum rectangle including all salient objects, which linearly speeds up the computation to 30 FPS. Some examples of optimal $\hat{\mathbf{v}}_{ki}$ under different display resolution are given in Fig. 4.

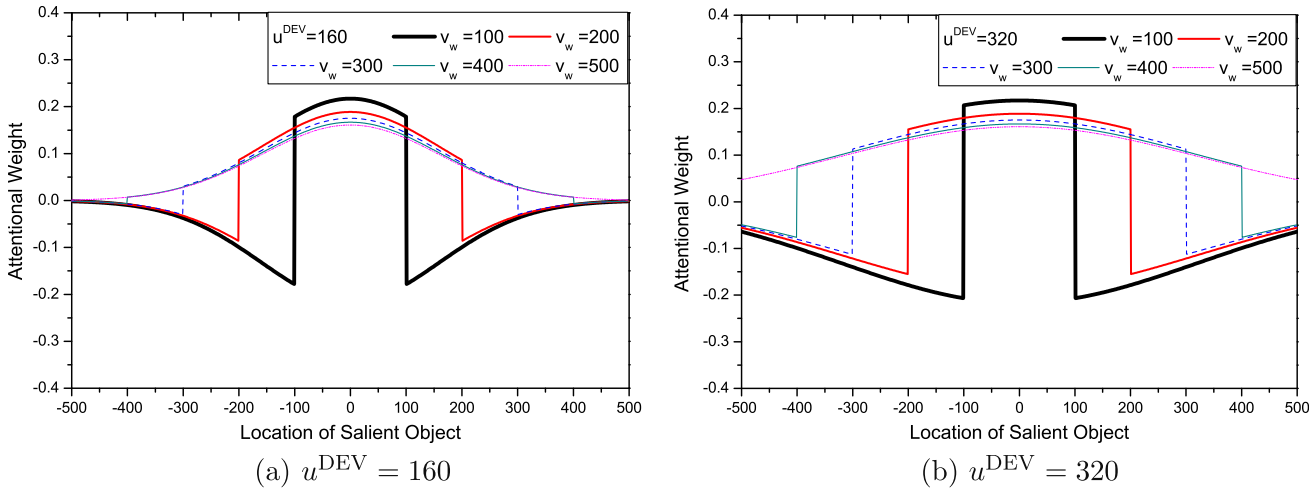
2.3.2. Selection of camera views for a given frame

Although we use data from multiple sensors, what really matters is not the number of sensors or their positioning, but the way we utilize those viewpoints to produce a unified virtual viewpoint which takes a good balance between local emphasis of details and global overview of scenarios. Since it is difficult to generate high-quality free viewpoint videos with the state-of-art methods, we only consider selecting a camera view from all presented cameras in the present work to make our system more generic. We define $\mathbf{c} = \{c_i\}$ as a camera sequence, where c_i denotes the camera index for the ith frame. A trivial understanding in evaluating a camera view is that all salient objects should be clearly rendered with few occlusions and high resolution. For the ith frame in the kth camera, we define occlusion rate of salient objects as the normalized ratio of the united area of salient objects with respect to the sum of their individual area, i.e.,

$$\mathscr{R}^{\mathrm{OC}}_{ki}(\mathbf{v}_{ki}) = \frac{\mathscr{N}_{ki}(\mathbf{v}_{ki})}{\mathscr{N}_{ki}(\mathbf{v}_{ki})-1}\left\{1 - \frac{\mathscr{A}\left[\bigcup_m (\mathbf{o}_{kim} \cap \mathbf{v}_{ki})\right]}{\sum_m \mathscr{A}[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}]}\right\}$$

where $\bigcup_m x_m$ calculates the union of all bounding boxes $\{x_m\}$. We use $\mathscr{N}_{ki}(\mathbf{v}_{ki}) = \sum_{m,\mathbf{o}_{kim}\cap\mathbf{v}_{ki}\neq\phi} 1$ to represent the number of visible objects inside viewpoint $\mathbf{v}_{ci}$. To normalize the occlusion ratio against various numbers of salient objects in different frames, we rescale $\mathscr{R}^{\mathrm{OC}}_{ki}(\mathbf{v}_{ki})$ into the range of 0–1 by applying $\mathscr{N}_{ki}(\mathbf{v}_{ki})/(\mathscr{N}_{ki}(\mathbf{v}_{ki})-1)$. We define the closeness of salient objects as average pixel areas used for rendering objects, i.e.,

$$\mathscr{R}^{\mathrm{CL}}_{ki}(\mathbf{v}_{ki}) = \log\frac{1}{\mathscr{N}_{ki}(\mathbf{v}_{ki})}\sum_m \mathscr{A}[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}]. \tag{8}$$



(a) $u^{\mathrm{DEV}} = 160$ (b) $u^{\mathrm{DEV}} = 320$

Fig. 3. Attentional weighting function proposed in the present work. We trade-off between closeness and completeness by controlling the tail and the peak sharpness of the weighting function.

Fig. 4. Behavior of viewpoint selection under different display sizes.

Also we define the completeness of this camera view as the percentage of included salient objects, i.e.,

$$\mathscr{R}_{ki}^{CP}(\mathbf{v}_{ki},\mathbf{u}) = \frac{1}{\sum_m \mathscr{I}(\mathbf{o}_{kim}|\mathbf{u})} \sum_{\substack{m \\ \mathbf{o}_{cim}\cap\mathbf{v}_{ci}\neq\phi}} \mathscr{I}(\mathbf{o}_{kim}|\mathbf{u}). \tag{9}$$

Accordingly, the interest gain of choosing the kth camera for the ith frame is evaluated by $\mathscr{I}_i(k|\mathbf{v}_{ki},\mathbf{u})$, which reads,

$$\mathscr{I}_i(k|\mathbf{v}_{ki},\mathbf{u}) = w_k(\mathbf{u})\mathscr{R}_{ki}^{CL}(\mathbf{v}_{ki})\mathscr{R}_{ki}^{CP}(\mathbf{v}_{ki},\mathbf{u})\exp\left[-\frac{\mathscr{R}_{ki}^{OC}(\mathbf{v}_{ki})^2}{2}\right]. \tag{10}$$

We weights the support of user preference on camera u^{CAM} to camera k by $w_k(\mathbf{u})$, which assigns a higher value to camera k if it is specified by the user and assigns a lower value if it is not specified. We then define the probability of taking the kth camera for the ith frame under $\{\mathbf{v}_{ki}\}$ as

$$P(c_i = k|\mathbf{v}_{ki},\mathbf{u}) \equiv \frac{\mathscr{I}_i(k|\mathbf{v}_{ki},\mathbf{u})}{\sum_j \mathscr{I}_i(j|\mathbf{v}_{ki},\mathbf{u})} \tag{11}$$

2.4. Generation of smooth viewpoint/camera sequences

A video sequence with individually optimized viewpoints will have obvious fluctuations, which leads to uncomfortable visual artifacts. We solve this problem by generating a smooth moving sequence of both cameras and viewpoints based on their individual optima. We use a graph in Fig. 5 to explain this estimation procedure, which covers two steps of the whole system, i.e., camera-wise smoothing of viewpoint movements and generation of a smooth camera sequence based on determined viewpoints. At first, we take $\hat{\mathbf{v}}_{ki}$ as observed data and assume that they are noise-distorted outputs of some underlying smooth results $\mathbf{v}_{ki}$. We use statistical inference to recover one smooth viewpoint sequence for each camera. Taking camera-gains of those derived viewpoints into consideration, we then generate a smooth camera sequence.

2.4.1. Camera-wise smoothing of viewpoint movement

We start from the smoothness of viewpoint movement on a video from the same camera. There are two contradictory strengths that drive the optimization of viewpoint movement: on one hand, optimized viewpoints should be closer to the optimal viewpoint of each individual frame; on the other hand, inter-frame smoothness of viewpoints prevents dramatic switching from occurring. Accordingly, we model smooth viewpoint movement as a Gaussian Markov Random Field (MRF), where the camera-wise smoothness is modeled as the priori of viewpoint configuration, i.e.,

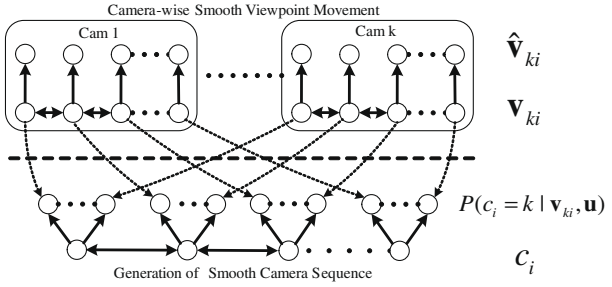


Fig. 5. Graph model for two-step estimation of smooth viewpoint movement.

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u}) = \frac{\exp\left(-\frac{1}{2}\sum_i\sum_{j\in\mathscr{N}_i}\mathscr{H}^{PRI}\right)}{\sum_{\{\mathbf{v}_{ki}\}}\exp\left(-\frac{1}{2}\sum_i\sum_{j\in\mathscr{N}_i}\mathscr{H}^{PRI}\right)}, \tag{12}$$

$$\mathscr{H}_{ij}^{PRI} = \frac{(v_{kix}-v_{kjx})^2}{2\sigma_{1x}^2} + \frac{(v_{kiy}-v_{kjy})^2}{2\sigma_{1y}^2} + \frac{(v_{kiw}-v_{kjw})^2}{2\sigma_{1w}^2}, \tag{13}$$

where $\mathscr{N}_i$ is the neighborhood of the ith frame, while a conditional distribution

$$P(\{\hat{\mathbf{v}}_{ki}\}|\mathbf{u},\{\mathbf{v}_{ki}\}) = \prod_i \frac{\exp\left(-\mathscr{H}_i^{LL}\right)}{\sum_{\mathbf{v}_{ki}}\exp\left(-\mathscr{H}_i^{LL}\right)} \tag{14}$$

$$\mathscr{H}_i^{LL} = \frac{(v_{kix}-\hat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} + \frac{(v_{kiy}-\hat{v}_{kiy})^2}{2\beta_{ki}\sigma_{2y}^2} + \frac{(v_{kiw}-\hat{v}_{kiw})^2}{2\beta_{ki}\sigma_{2w}^2} \tag{15}$$

describes the noise that produces the final results. We add a parameter β_{ki} to control the flexibility of current frame in smoothing. A smaller β_{ki} can be set to increase the tendency of current frame in approaching its locally optimal viewpoint. Estimation of optimal viewpoints $\{\mathbf{v}_{ki}\}$ is done by maximizing the posterior probability of $\{\mathbf{v}_{ki}\}$ over observed $\{\hat{\mathbf{v}}_{ki}\}$, i.e., $P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$, which is expressed by a Gibbs canonical distribution [21], i.e.,

$$P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) = \frac{\exp\{-\mathscr{H}^V\}}{\sum_{\{\mathbf{v}_{ki}\}}\exp\{-\mathscr{H}^V\}} \tag{16}$$

with

$$\mathscr{H}^V = \frac{1}{2}\sum_i\sum_{j\in\mathscr{N}_i}\mathscr{H}^{PRI} + \sum_i \mathscr{H}_i^{LL}. \tag{17}$$

In statistical physics, the optimal configuration of largest posterior probability is determined by minimizing the following free energy [22]:

$$\mathscr{F}^V = \langle\mathscr{H}^V\rangle - \langle\ln P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})\rangle \tag{18}$$

where $\langle x\rangle = \sum_{\{\mathbf{v}_{ki}\}} xP(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ is the expectation value of a quantity x. We then form the following criterion by considering the normalization constraint of $P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ as:

$$\mathscr{L}^V = \mathscr{F}^V + \eta\left(1 - \sum_{\{\mathbf{v}_{ki}\}} P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})\right) \tag{19}$$

where η is a Lagrangian multiplier. We use the Mean-field approximation [22] which assumes that $P(\{\mathbf{v}_{ki}\}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) \approx \prod_i P(\mathbf{v}_{ki}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ to decouple two-body correlations. By taking differential of $\mathscr{L}^{VP}$ with respect to $P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ and setting it to zero, we obtain the optimal estimation for $P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})$ as

$$\begin{aligned} 0 &= \frac{\partial\mathscr{L}^{VP}}{\partial P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\})} \\ &= \sum_{j\in\mathscr{N}_i}\frac{v_{kix}^2 - 2v_{kix}\langle v_{kjx}\rangle^T}{2\sigma_{1x}^2} + \frac{(v_{kix}-\hat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2} + \ln P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) - \eta. \end{aligned} \tag{20}$$

Therefore, we have posterior probability

$$P(v_{kix}|\mathbf{u},\{\hat{\mathbf{v}}_{ki}\}) \propto \exp\left\{-\sum_{j\in\mathcal{N}_i}\frac{(v_{kix}-\langle v_{kjx}\rangle)^2}{2\sigma_{1x}^2}-\frac{(v_{kix}-\hat{v}_{kix})^2}{2\beta_{ki}\sigma_{2x}^2}\right\}. \tag{21}$$

Since it is a Gaussian distribution whose mean value has the maximized probability, the optimal viewpoint for v_{kix} is solved as

$$v_{kix}^* = \langle v_{kix}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2x}^2\beta_{ki}\langle v_{kjx}\rangle + \hat{v}_{kix}\sigma_{1x}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2x}^2\beta_{ki}+\sigma_{1x}^2}, \tag{22}$$

$$v_{kiy}^* = \langle v_{kiy}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2y}^2\beta_{ki}\langle v_{kjy}\rangle + \hat{v}_{kiy}\sigma_{1y}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2y}^2\beta_{ki}+\sigma_{1y}^2}, \tag{23}$$

$$v_{kiw}^* = \langle v_{kiw}\rangle = \frac{\sum_{j\in\mathcal{N}_i}\sigma_{2w}^2\beta_{ki}\langle v_{kjw}\rangle + \hat{v}_{kiw}\sigma_{1w}^2}{\sum_{j\in\mathcal{N}_i}\sigma_{2w}^2\beta_{ki}+\sigma_{1w}^2} \tag{24}$$

with optimal results for v_{kiy} and v_{kiw} also given by similar derivation. We use $\mathbf{v}_{ki}^*$ in following sections to denote the optimal viewpoint represented by v_{kix}^*, v_{kiy}^*, and v_{kiw}^*.

2.4.2. Smoothing of camera sequence

A smooth camera sequence will be generated from determined viewpoints. For simplicity, we use $\rho_{ki} \equiv \log P(c_i = k|\mathbf{v}_{ki}^*,\mathbf{u})$ to shorten the formulation, which is computed by using Eq. (11). We have to trade-off between minimizing camera switching and maximizing the overall gain of cameras. We use another MRF to model these two kinds of strengths. The smoothness of camera sequence is modeled by a Gibbs canonical distribution, which reads,

$$P(\{c_i\}|\{\mathbf{v}_{ki}^*\},\mathbf{u}) = \frac{\exp(-\mathscr{H}^C)}{\sum_{\{c_i\}}\exp(-\mathscr{H}^C)}, \tag{25}$$

with

$$\mathscr{H}^C = -\gamma\sum_{i,k}\delta_{c_i,k}\rho_{ki} - \frac{(1-\gamma)}{2}\sum_i\sum_{j\in\mathcal{N}_i}\alpha_{ij}\delta_{c_i,c_j}. \tag{26}$$

where α_{ij} is a parameter to normalize the relative strength of smoothing with respect to the size of neighborhood, which reads

$$\alpha_{ij} = \frac{K}{|j-i|\sum_{l\in\mathcal{N}_i}\frac{1}{|l-i|}}. \tag{27}$$

γ is a hyper-parameter for controlling the smoothing strength. We use Mean-field approximation again to achieve the optimal estimation, which reads that $P(\{c_i\}|\{\mathbf{v}_{ki}^*\},\mathbf{u}) \approx \prod_i P(c_i|\{\mathbf{v}_{ki}^*\},\mathbf{u})$. We omit the detailed derivation and only show the final result, which derives that the marginal probability of taking camera k for the ith frame is

$$\begin{aligned} P(c_i = k|\{\mathbf{v}_{ki}^*\},\mathbf{u}) &= \langle\delta_{c_i,k}\rangle^C \\ &= \frac{\exp\left\{(1-\gamma)\sum_{j\in\mathcal{N}_i}\alpha_{ij}\langle\delta_{c_j,k}\rangle^C + \gamma\rho_{ki}\right\}}{\sum_k\exp\left\{(1-\gamma)\sum_{j\in\mathcal{N}_i}\alpha_{ij}\left\langle\delta_{c_j,k}\right\rangle^C + \gamma\rho_{ki}\right\}}, \end{aligned} \tag{28}$$

where $\langle x\rangle^C = \sum_{\{c_i\}} xP(\{c_i\}|\{\mathbf{v}_{ki}^*\},\mathbf{u})$ is the expectation value of a quantity x. The smoothing process is performed by iterating the following fixed-point rule until reaching convergence:

$$\langle\delta_{c_i,k}\rangle^C = \frac{\exp\left\{(1-\gamma)um_{j\in\mathcal{N}_i}\alpha_{ij}\langle\delta_{c_j,k}\rangle^C + \gamma\rho_{ki}\right\}}{\sum_k\exp\left\{(1-\gamma)\sum_{j\in\mathcal{N}_i}\alpha_{ij}\langle\delta_{c_j,k}\rangle^C + \gamma\rho_{ki}\right\}}. \tag{29}$$

After convergence, we select the camera which maximizes $\langle\delta_{ci,k}\rangle^C$, i.e.,

$$c_i^* = \arg\max_{c_i}\langle\delta_{c_i,k}\rangle^C. \tag{30}$$

3. Experimental results and discussions

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 [1] and more detailed explanation about the acquisition settings could be found in Ref. [23]. Briefly, those cameras are all Arecont Vision AV2100M IP cameras, whose positions in the basketball court are shown in Fig. 6. The fish-eye lenses used for the top-view cameras are Fujinon FE185C086HA-1 lenses. Frames from seven cameras were all sent to a server, where the arrival time of each frame was used in synchronizing different cameras. In Fig. 7, samples images from all the seven cameras are given. Due to the limited number of cameras, we set most of the cameras to cover the left court. As a result, we will mainly focus on the left court to investigate the performance of our system in personalized production of sports videos.

In Section 3.1, we use manually annotated meta-data as ground-truth to investigate the behavior of the system under different parameters. We then test the robustness of our system against incomplete meta-data in Section 3.2. In Section 3.3, positions are computed automatically using Ref. [18]. Clock signals are monitored to segment the sequence into shots that are independently produced.

Since video production still lacks an objective rule for performance evaluation, many parameters are heuristically determined based on subjective evaluation. We defined several salient objects and relationship between object type and interest is given in Table 1. If the user shows special interests on one salient object, the weight will be multiplied by a factor 1.2. For viewpoint smoothing, we set all β_{ki} to 1 for camera-wise viewpoint smoothing in the following experiments. We also let $\sigma_{1x} = \sigma_{1y} = \sigma_{1w} = \sigma_1$ and $\sigma_{2x} = \sigma_{2y} = \sigma_{2w} = \sigma_2$.

3.1. Behavior of our system

We first investigate the behavior of our system under different parameters to verify their efficiency and the correctness of implementation. A short video clip with about 1200 frames is used to demonstrate behavioral characteristics of our system, especially its adaptivity under limited display resolution. This clip covers three ball-possession periods and includes five events in total. In Fig. 8, we show time spans of all events, whose most highlighted moments are also marked out by red solid lines. We will explore the efficiency of each individual processing step of our method,

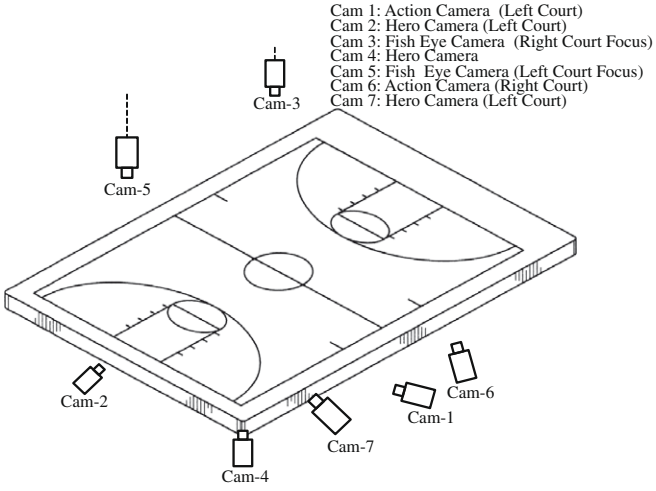


Fig. 6. Camera plans in gathering experimental video data.

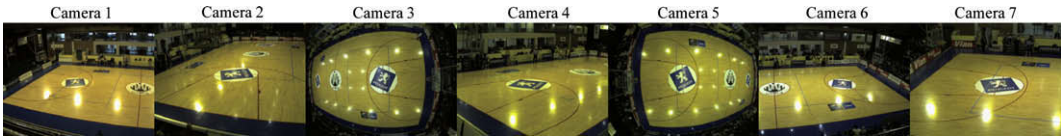


Fig. 7. Sample views gathered by different cameras.

and then make an overall evaluation based on finally generated outputs. Due to the page limitation, numerical results are given and depicted by graphs in the present paper while their corresponding videos are only available in the website of APIDIS project [1]. Reviewers are invited to download video samples produced based on different user preferences to subjectively evaluate the efficiency and relevance of the proposed approach.

We start from investigating the performance of our method for individual selection of viewpoints. Camera-wise sequences of automatically determined viewpoints by our method are shown in Fig. 9, where widths of optimal viewpoints under three different display resolutions, i.e., 160×120, 320×240, and 640×480 are displayed for all the seven cameras. Weak viewpoint smoothing has been applied to improve the readability of generated videos, where the smoothing strength is set to $\sigma_2/\sigma_1 = 4$. From comparison of results under three different display resolutions, the most obvious finding is that a higher display resolution leads to a larger viewpoint width while a lower display resolution prefers a smaller viewpoint size, just as we have expected from our selection criterion. Since cameras 1, 6, and 7 only cover half the court, their sizes of viewpoints will be fixed when all players are in the other half court, which explains the flat segments in their corresponding sub-graphs. From the video data, we could further confirm that even when the display resolution is very low, our system will extract a viewpoint of a reasonable size where the ball is scaled to a visible size. Although in some frames only the ball is displayed for the lowest display resolution, it will not cause a problem because those frames will be filtered out by post camera selection.

Viewpoint sizes of smoothed sequences under different smoothing strengths are compared in Fig. 10a. With all other parameters being the same, the ratio of σ_2 to σ_1 is tuned for all the five cases. A higher ratio of σ_2 to σ_1 corresponds to a stronger smoothing process while a smaller ratio means weaker smoothing. When $\sigma_2/\sigma_1 = 1$ where very weak smoothing is applied, we obtain a quite accidented sequence, which results in a flickering video with a lot of dramatic viewpoint movements. With the increasing of σ_2/σ_1 ratio, the curve of viewpoint movement becomes to have less sharp peaks, which provides perceptually more comfortable contents. Another important observation is that generated sequences will be quite different from our initial selection based on saliency information, if too strong smoothing has been performed with a very large σ_2/σ_1. This will cause such problems as the favorite player or the ball is out of the smoothed viewpoint. Ratio σ_2/σ_1 should be determined by considering the trade-off between locally optimized viewpoints and globally smoothed viewpoint sequences. By visually checking the generated videos, we consider that results with a weak smoothing such as $\sigma_2/\sigma_1 = 4$ are already perceptually acceptable by viewing the demo video.

We then verify our smoothing algorithm for camera sequence. Smoothed camera sequences under various smoothing strength γ are depicted in Fig. 10b. The smoothing process takes the probability defined in Eq. (11) as initial values, and iterates the fixed-point updating rule with a neighborhood of size thirty until convergence. A camera sequence without smoothing corresponds to the topmost sub-graph in Fig. 10b, while the sequence with the strongest smoothing is plotted in the bottom sub-graph. It is clear that there are many dramatic camera switches in an unsmoothed sequence, which leads to even more annoying visual artifacts than fluctuated viewpoint position, as we can see from the generated videos. Therefore, we prefer strong smoothing on camera sequences and will use $\gamma = 0.8$ in following experiments.

Table 1
Weighting of different salient objects.

Object type	Ball	Player	Judge	Basket	Coach bench	Others
$\mathscr{I}(\mathbf{o}_{kim}\|\mathbf{u})$	2	1	0.8	0.6	0.4	0.2

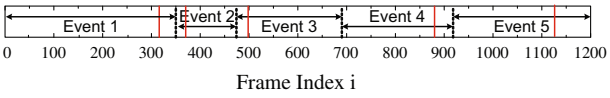


Fig. 8. A short video clip with 1200 frames is used to demonstrate the system. There are five clock-events inside this clip. For each event, the most highlighted moment is marked in a red solid bar. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 11a and b, we compare viewpoints and cameras in generated sequences with respect to different display resolutions, respectively. From top to bottom, we show results for display resolution $u^{\mathrm{DEV}} = 160,\ 320$, and 640 in three sub-graphs. When the same camera is selected, we observe that a larger viewpoint is preferred by a higher display resolution. When different cameras are selected, we need to consider both the position of selected camera and the position of determined viewpoint in evaluating the coverage of output scene. Again, we confirm that sizes of viewpoints increase when display resolution becomes larger. Before the 400th frame, the event occurs in the right court. We find that the 3rd camera, i.e., the top-view with wide-angle lens, appears more often in the sequence of $u^{\mathrm{DEV}} = 640$ than that of $u^{\mathrm{DEV}} = 160$ and their viewpoints are also broader, which proves that a larger resolution prefers a wider view. Although the 2nd camera appears quite often in $u^{\mathrm{DEV}} = 160$, its corresponding viewpoints are much smaller in width. This camera is selected because it provides a side view of the right court with salient object gathered closer than other camera views due to projective geometry. For the same reason, the 3rd camera appears more often in $u^{\mathrm{DEV}} = 160$ when the game moves to the left court from the 450th frame to the 950th frame. This conclusion is further confirmed by thumbnails in Fig. 12, where frames from index 100 to 900 are arranged into a table for the above three display resolutions.

Due to the fact that different cameras were selected, viewpoints determined under $u^{\mathrm{DEV}} = 640$ seem to be closer than those under $u^{\mathrm{DEV}} = 320$ in the last five columns of Fig. 12. This reflects the inconsistency of relative importance between completeness and closeness in viewpoint selection. Since only central points of salient objects are calculated in the criteria for viewpoint selection, result viewpoints are not continuous under different resolution. Although camera-7 is similar to camera-1 with linear zooming in, their optimal viewpoints might have different emphases on completeness and closeness. This consistency also exists in separated selection of cameras and viewpoints. If viewpoint selection focuses more on closeness and camera selection focuses more on completeness, a small cropping area on camera-7 will be first selected in viewpoint selection for $u^{\mathrm{DEV}} = 320$, and then be rejected in the

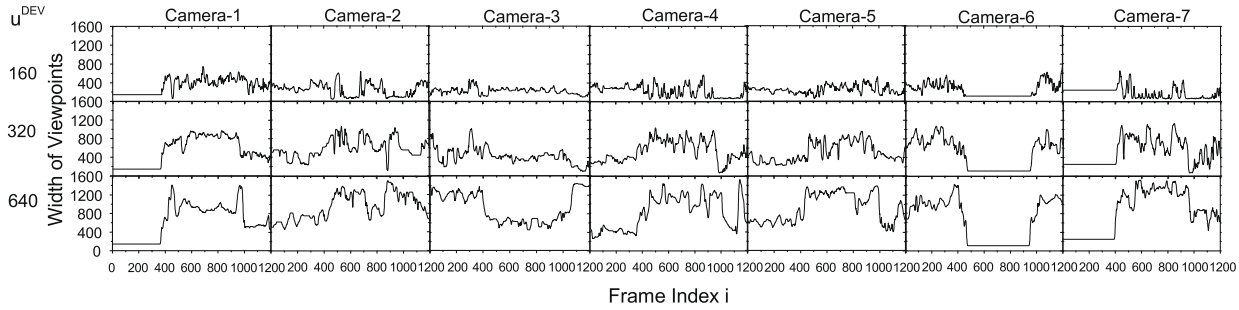


Fig. 9. Camera-wise viewpoint sequences automatically selected under three different display resolutions. Weak viewpoint smoothing has been applied and the smoothing strength is set to $\sigma_2/\sigma_1 = 4$.

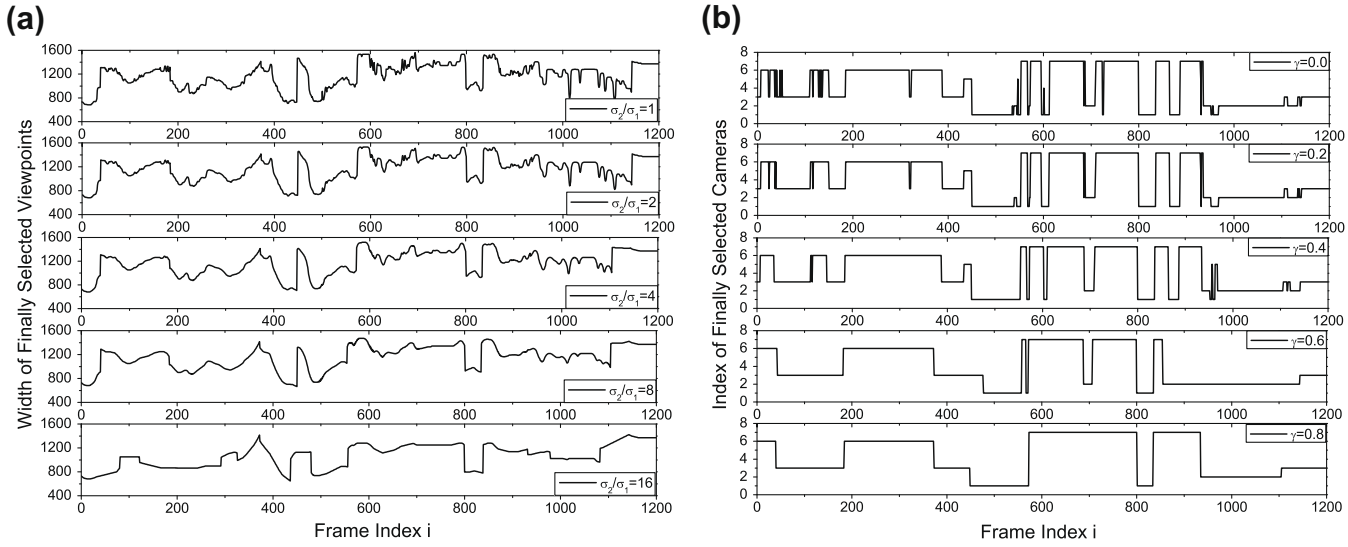


Fig. 10. Behavior of optimized camera/viewpoint sequence under different smoothing strengths. Display resolution is set to $u^{DEV} = 640$.

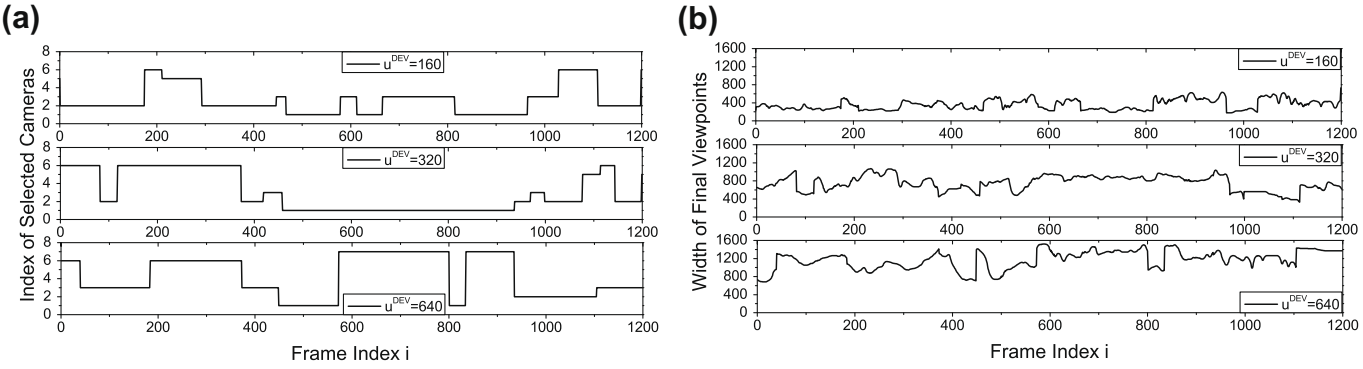


Fig. 11. Comparison of generated camera and viewpoint sequences under three different display resolutions 160, 320 and 640. Viewpoint smoothing strength is $\sigma_2/\sigma_1 = 4$ and camera smoothing strength γ is set to 0.8.

following camera selection due to insufficient completeness. Subjective test will help us to tune the relative weighting of completeness and closeness. It is more important to implement simultaneous selection of viewpoints and cameras, which requires both inclusion of positional information of cameras such as using homography, and an analytically solvable criterion for viewpoint selection. These issues are our major work in the near future.

In all above experiments, no narrative user preferences are included. If the user has special interests on a certain camera view, we could assign a higher weighting $w_k(\mathbf{u})$ to the specified camera. In our case, we set $w_k(\mathbf{u}) = 1.0$ for not-specified cameras and $w_k(\mathbf{u}) = 1.2$ for a user-specified camera. We compare the camera sequences under different preferences in Fig. 13. As we can easily see from the graph, a camera appears more times when it is specified, which reflects the user preference on camera views. As for user preference on teams or players, the difference between viewpoints with and without user preferences is difficult to tell without a well-defined evaluation rule, because all players are always cluttered together during the game. In fact, we are more interested in reflecting user preferences on players or teams by extracting their

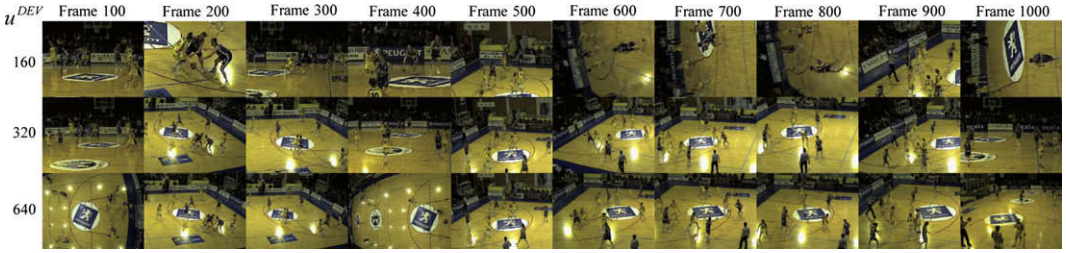


Fig. 12. Thumbnails of frames in generated sequences under three different display resolutions 160, 320 and 640. Viewpoint smoothing strength is $\sigma_2/\sigma_1 = 4$ and camera smoothing strength γ is set to 0.8.

relative frames. We thus omit the results on player or team selection, but explore them later along with results from our future work on video summarization.

3.2. Robustness of our system against incomplete meta-data

To verify our system in real application scenario, we have to investigate the robustness of our system against incomplete meta-data. Starting from manually annotated salient objects, we performed random deletion of salient objects, including both players and referees, to simulate the error of automatic detection, where each salient object was deleted according to a specified probability. We compared the production result based on complete meta-data to five incomplete cases, with different deletion probabilities from 10% to 50%, and plotted the width of selected viewpoints after weak smoothing ($\sigma_2/\sigma_1 = 1$) in Fig. 14. Four cameras are plotted, which include camera-1 for the side-view of the left court, camera-6 for the side-view of the right court, camera-3 for the top-view of the right court and camera-5 for the top-view of the left court. Our system produced more noisy sequences of selected viewpoints when probabilities of object deletion went higher. However, the overall profile is still preserved in the noisy case. We also plot the results after stronger smoothing, with $\sigma_2/\sigma_1 = 8$ in Fig. 15. After performing the smoothing process of viewpoint sequences, these profiles of selected viewpoints from incomplete meta-data became quite closer to that from complete meta-data, although they were deteriorated due to the missing of some salient objects. Therefore, inclusion of smoothing process is not only meaningful for removing flickering artifacts, but also important in improving the robustness against incompleteness of salient objects.

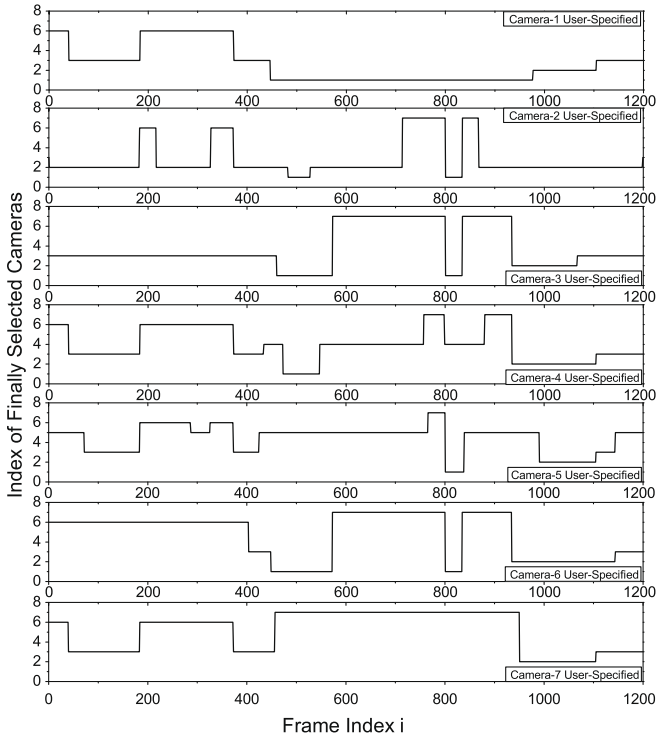


Fig. 13. Comparison of generated camera sequences under different user preferences on camera views for $u^{DEV} = 640$. Viewpoint smoothing strength is $\sigma_2/\sigma_1 = 4$ and camera smoothing strength γ is set to 0.8.

Although we have achieved some promising progress in automatic ball detection, it is still a very difficult task because the ball is small, similar to the background and sometimes occluded by the player. Instead of using automatically detected ball positions, we investigate the robustness of our approach against missing of ball positions in the present paper, by comparing two videos produced before and after removing all positional information of the ball in Fig. 16. Since the ball is usually inside the major player cluster, missing of ball will not cause significant difference for most of the time. Significant differences occurred within the period from the 1000th to 1200th frames, when the ball was in the left court with two players and all other players were in the right court. The inclusion of ball positions is very important if we want to assure that the ball is always included in the view.

On the other hand, several dedicated player detectors with high detection rate have been successfully implemented for the APIDIS project [18–20]. In Fig. 17, we compare results based on manually annotated meta-data to those based on automatically detected salient objects obtained in Ref. [18], where accurate player detection is achieved by fusing the calibrated foreground masks from multiple views to solve the problem of reflection, occlusion and shadow in the single view case. In both cases, we use manually annotated ball positions. Significant difference appears again in the period from the 1100th to 1200th frame in camera-3 and camera-6, both focusing on the right court. By checking the video, all players and referees evenly scattered over the whole court during that period, with two players and one referee in the left court preparing for starting a new round and all other players and referees in the right court. Since automatically detected results have missing objects quite often in the right-bottom corner of the court, it leads to a smaller viewpoint than the one we got from manually annotated salient objects. When players gather together, misdetection of objects does not affect selected viewpoints much. Furthermore, we conclude that the automatic player detector is very accurate, by both viewing Fig. 17 and verifying their bounding boxes in the videos.

3.3. Subjective evaluation

Before designing our plan for subjective test, we presented a set of demo videos produced under different parameters to 25 viewers during WIAMIS 2009, and asked them to fill a questionnaire, which

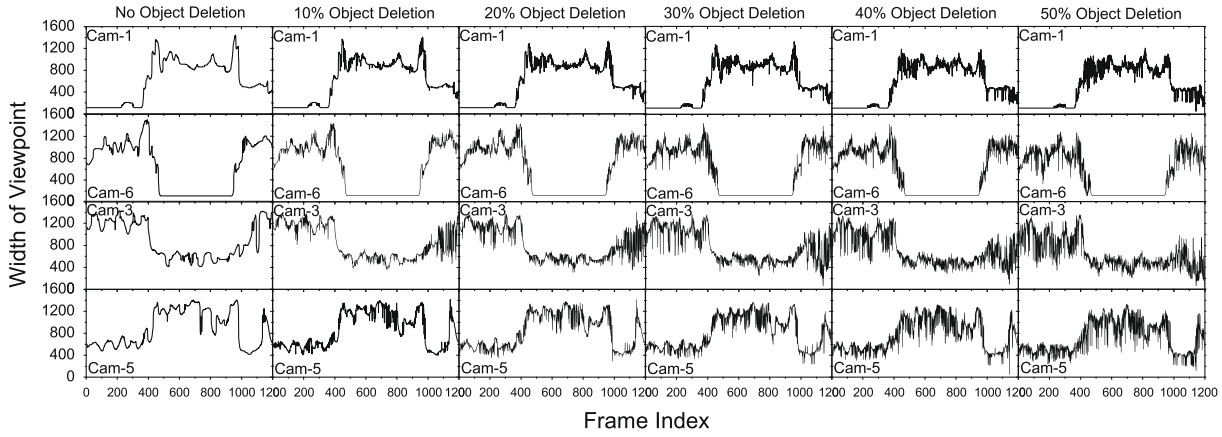


Fig. 14. The width of selected viewpoint is plotted for both complete and incomplete meta-data for four cameras after performing weak smoothing ($\sigma_2/\sigma_1 = 1$).

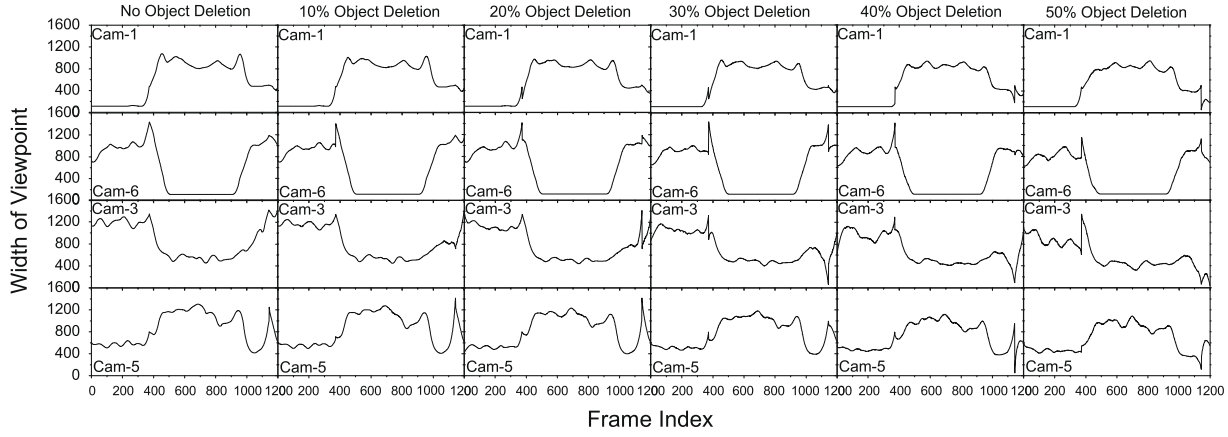


Fig. 15. The width of selected viewpoint is plotted for both complete and incomplete meta-data for four cameras after performing strong smoothing ($\sigma_2/\sigma_1 = 8$).

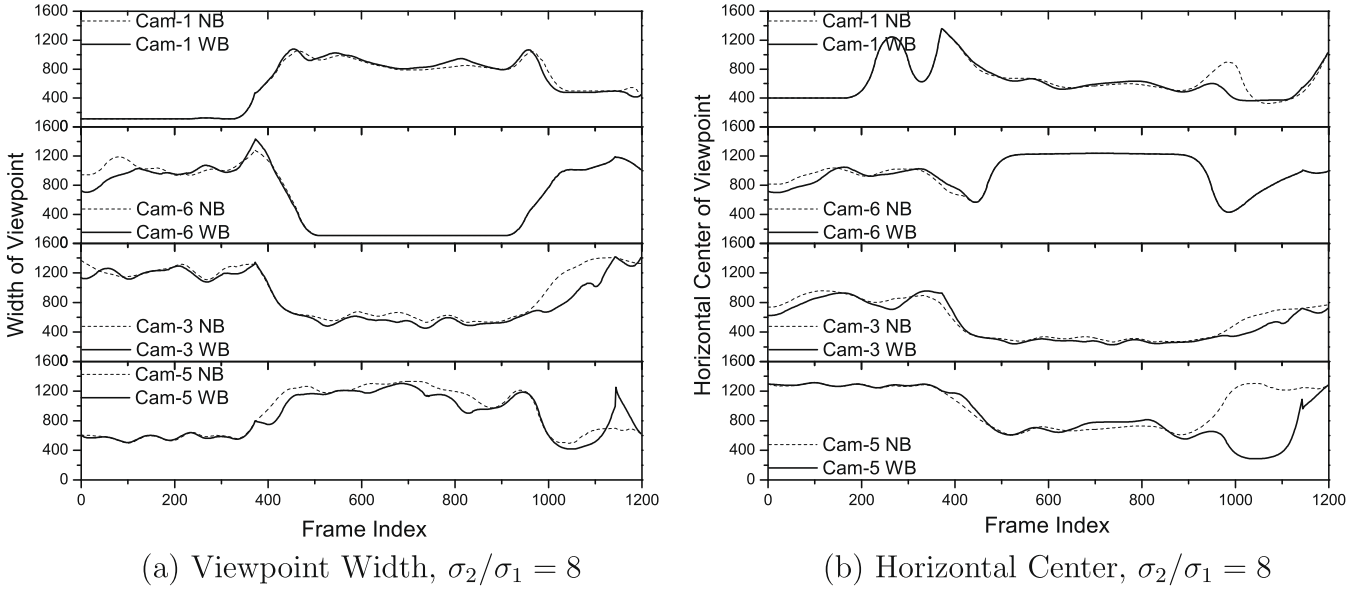
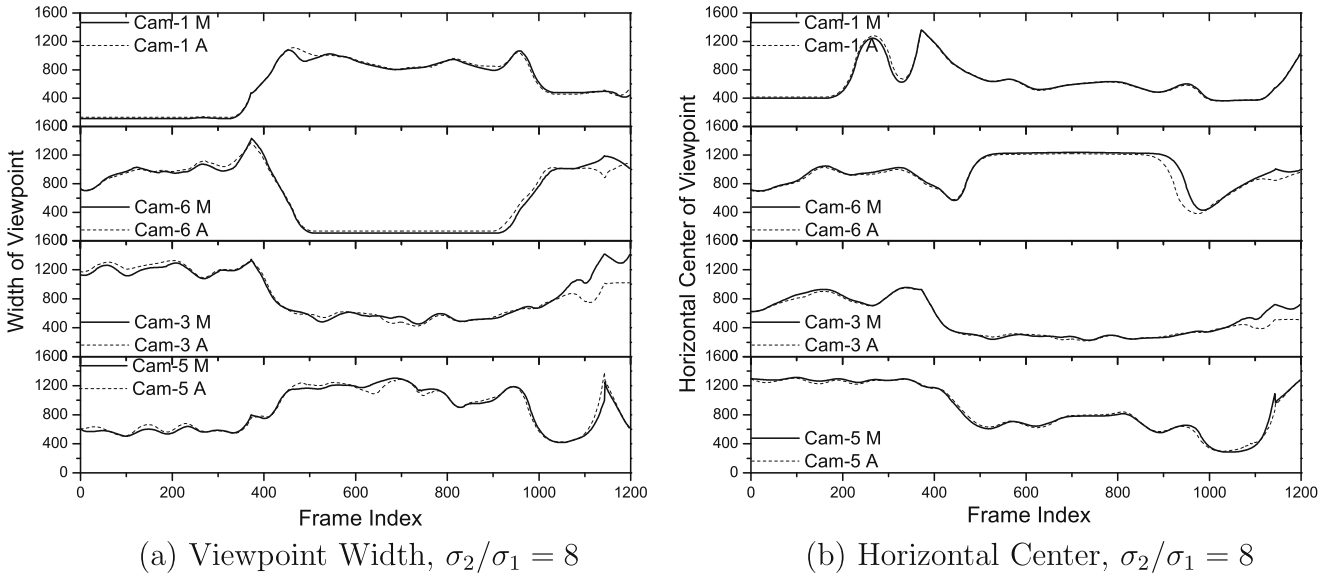


Fig. 16. Selected viewpoints are plotted for results from both annotated meta-data with and without ball positions.

consists of 10 questions. The corresponding answers are plotted in Fig. 18. They confirm that changing the behavior of the production system by tuning parameters in Section 3.1 is not only reflected in the above graphs, but also visually perceptible. The importance of our three key concepts (smoothness, closeness and completeness) and two important user preferences (preferred view and preferred player) is recognized by most of the viewers. Furthermore, nearly all viewers agree that the concept of autonomous production sys-



(a) Viewpoint Width, $\sigma_2/\sigma_1 = 8$ (b) Horizontal Center, $\sigma_2/\sigma_1 = 8$

Fig. 17. Selected viewpoints are plotted for results from both manually annotated meta-data and automatically detected salient objects.

tem is relevant. It is interesting to find that utilization of top view is not well accepted by the viewers, which is the major production artifact reported in the questionnaire. Unstable viewpoint movement and low video quality of zoomed viewpoints are two other important artifacts.

Based on the above results, we disabled top-view cameras and performed strong smoothing to stabilize the viewpoint movement when preparing test materials for subjective evaluation. Testing materials were produced from a longer period of over 18 min, which covers the whole second-quarter of the basketball match. Manually annotated ball positions, and automatically detected players and referees were collected [18]. We also collected two production results manually made by experts as referential materials. The expected device resolution was set to 640×480 and parameters for automatic production were $\sigma_2/\sigma_1 = 8$, $\gamma = 0.3$. We design the process of subjective evaluation as first dividing the second-quarter into shorter segments, randomly selecting a clip from these three producing strategies for each segment, and then displaying it to a viewer. This viewer is then asked to rate the current clip, without knowing its corresponding production strategy. Two distinct notions of rating have been considered, each of them leading to an experiment. In the first experiment, we divided the video into 33 segments and asked 20 viewers to rate their global appreciation on each segment. Although the automatically produced video received more "highest" scores than manually produced videos, most of the segments were evaluated as "low", as shown in Fig. 19a. Poor sensor quality during acquisition (IP video surveillance cameras), and significant deterioration of video quality after cropping, resampling and re-encoding might have distracted the viewers' attention from comparing the performance of various producing strategies to evaluating the image quality. Therefore, in a second experiment of subjective evaluation, we divided the video into 30 segments by merging some over-short segments, and the viewers were instructed to focus their evaluation on viewpoint/camera selection. We plotted results from 10 reviewers in Fig. 19b, where the difference between strategies becomes

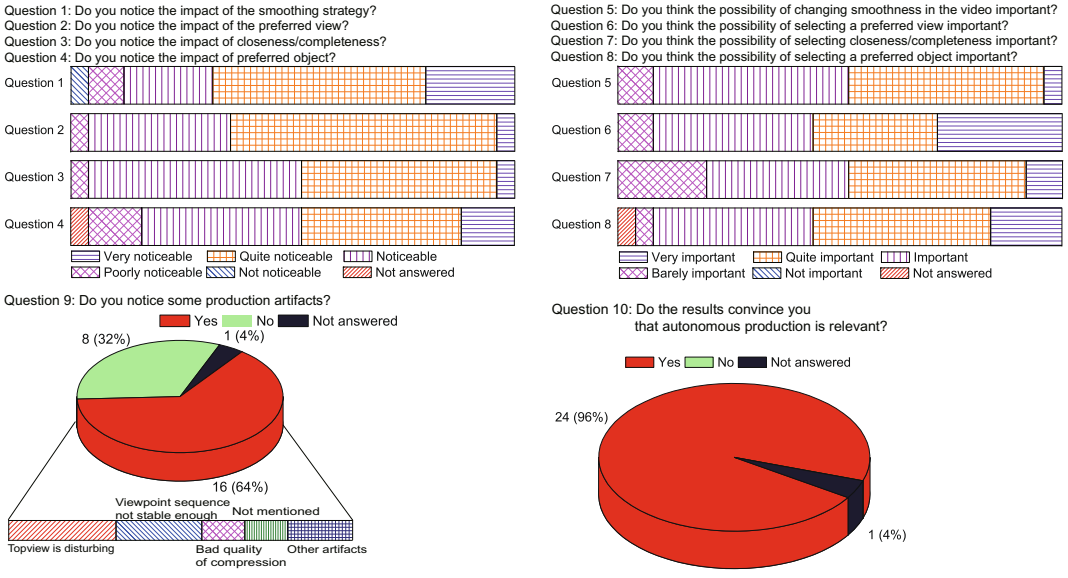


Fig. 18. Results of user questionnaires after presenting demo videos to 25 viewers in WIAMIS 2009.

more significant and automatically produced video received higher grades than other manual strategies in viewpoint/camera selection.

We further plot user-wise and segment-wise average scores of three different production strategies in the second round test along with their corresponding numbers of samples in Fig. 20a and b, respectively. The average score is computed by assigning 1 to "Unacceptable", 2 to "Acceptable", 3 to "Nice", and 4 to "Excellent". For a given segment, if one strategy was never selected due to our small number of viewers, we set the average score of this strategy on the current segment to zero, so that in Fig. 20b we can easily separate this invalid case from other valid data. The 10 viewers are further divided into two categories, i.e., two basketball coaches and eight general public, which are expected to have different viewing purposes. We summarize results of subjective evaluation from Figs. 19 and 20 as follows:

1. Our production strategy is evaluated to have the best performance, according to the fact that automatically produced video received higher grades than other manually produced results, no matter whether the viewers are instructed to focus on global quality or on viewpoint/camera selection in Fig. 19. This conclusion is also supported by the fact that automatically produced video is preferred by most of the viewers (Fig. 20a) on the majority of segments (Fig. 20b).
2. User appreciation varies with the viewing purpose, which makes automatic and personalized production of video important and relevant. In Fig. 20a, the first two users were basketball coaches, while all other eight viewers were all general public. We saw quite different trends on these two kinds of users, i.e., viewers from general public preferred automatic results, while the two coaches preferred expert results. By asking for further comments from these two coaches, we find that coaches have different requirements from common users. For example, they mainly care about stability of camera selection. They do not like camera switching during an action. For them, close viewpoint is an important, but secondary concern to continuous rendering of the action. By checking manually produced videos, we find that experts used two main cameras (camera-1 and camera-6) in 80–90% of the whole time and only resized the viewpoints in a few special cases, such as close-up view of players/referee during free-shoot actions. In order to reduce the burden of each viewer in subjective evaluation, we used only one automatically produced result which was targeted at satisfying general users, which is thus less suitable to the requirement of coaches. However, it is still much easier to meet different requirements from various users by changing parameters and including special rules in the automatic video production than by manually producing different versions of videos. For example, we can choose parameters in viewpoint selection to favor a wider viewpoint and disable camera switching within each action to further stable the scene, to meet the special requirement of coaches.
3. Since only 10 viewers have evaluated the videos, we only check segments where the significant difference between two production strategies is supported by at least three viewers on each strategy in Fig. 20b. Our automatic system is stronger in producing a closer view sequence with less occlusions and smoother camera switching, which is reflected by segments 3, 13, 15 and 30. On the contrary, manual production is better in dealing with complicated scenarios by fine manipulation of camera switching. For example, to present the free-shoot actions in segments 12, 22 and 28, both experts (especially expert-2) use a series of camera views, including both close-up views on dominant objects (referee and player) and top-views of the ball trail, which we think might be better in scene organization than our automatic system because they adopted conventional principles widely used in the field of TV production. We explain the difference in subjective appreciation as follows. On the one hand, manual results receive lower grades than automatic results mainly due to the fact that the automatic system stays closer to the action, without loosing it. On the other hand, the weakness of our automatic system is obvious in dealing with the long break in segment 22, where expert results focus the viewpoint on one of the team during the break while our automatically produced video focuses on the referees who were holding the ball. Hence, in the future, we plan to study the necessity of integrating and mimicking some conventional production principles adopted in some specific game actions (e.g. free-throw, dead time, etc.) in our system.

4. Related works

User-centric adaptation of video contents is required in our evolving information environment to meet requirements of various users using different transmission methods and heterogeneous

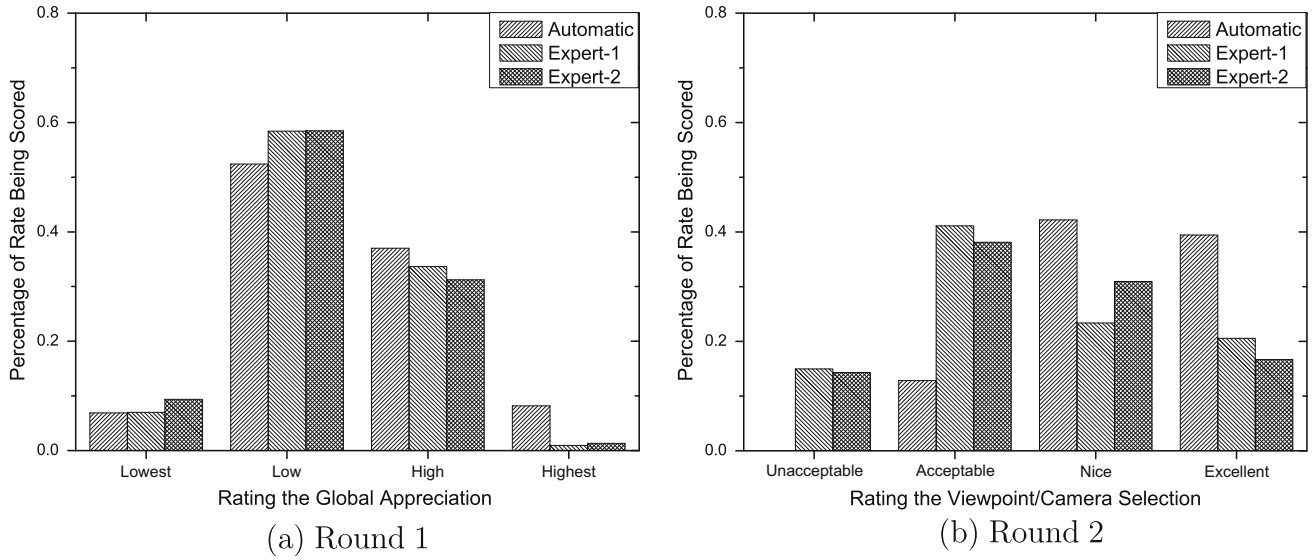


Fig. 19. Results of subjective evaluation between automatically produced video, and two videos that have been manually produced by experts.

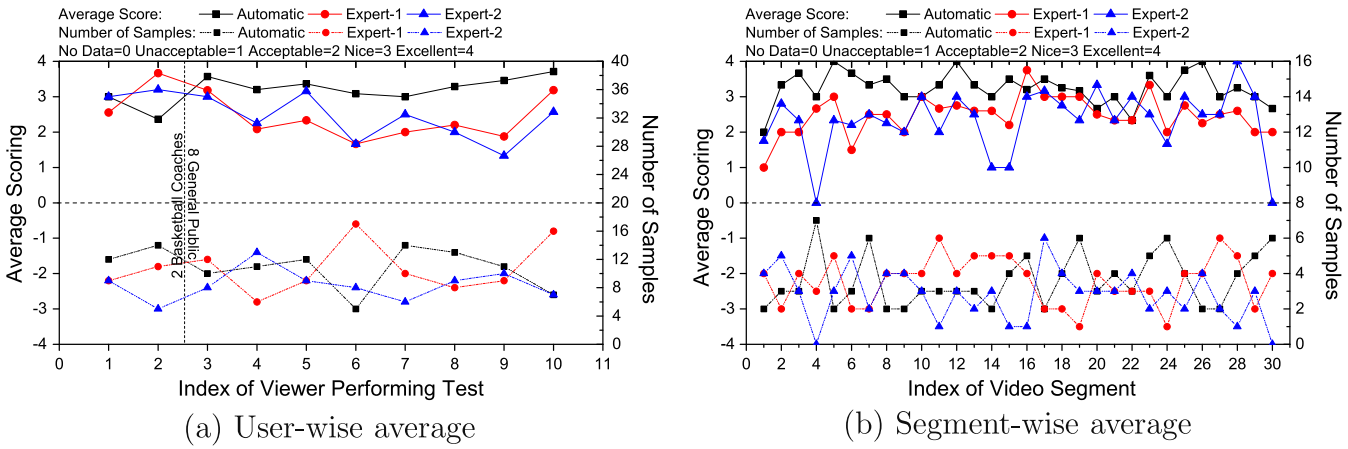


Fig. 20. Results on user-wise and segment-wise average scoring of three different production strategies evaluated by viewers in the second subjective test. The 10 viewers are further divided into two categories, i.e., two basketball coaches and eight general public.

terminal devices. Compared to video summarization for frame selection and interactive video streaming for adaptive video compression and transmission, automatic editing of frame contents is a less studied topic in the whole working flow of video generation, maybe due to the difficulty of analyzing video contents.

In the literature, previous approaches proposed for autonomous multi-camera video production can be classified into two major categories. Methods in the first category only perform camera selection, where camera switching is triggered by certain detected activities, such as an object entering the field of view or an audio event happening. Ref. [25] proposed a system for multimodal meetings, which selects the camera based on pre-defined rules, such as focusing more on the person who speaks first and longer, avoiding fast camera switching and suppressing too long shots. Ref. [24], also designed for conference, recognizes view types from pre-recorded conference/lectures and re-organizes a brief version of the contents by selecting the camera view. Camera selection for athletic videos is discussed in Ref. [26] based on rules defined on user preferences and the characteristics of athletic events. The second category captures a rich visual signal, either based on omni-directional cameras or on wide-angle multi-camera setting, to offer some flexibility in rendering the scene to the receiver-end. The approach proposed in Ref. [27] allows both user-control and computer control of selecting the view area from an omni-directional capturing system by automatic detection of specified activity in the scenario of meeting room, based on two simple rules, i.e., to switch to the new speaker and to avoid too long shots on the same person. In Refs. [28,29], a best shot is selected from a list of candidate shots for each scene from a video conference and a multiplayer game tv show, according to a pre-defined set of cenematic rules. However, in Refs. [27–29], the shot parameters (i.e., the cropping parameters in the view at hand) stay fixed until the camera is switched. Furthermore, a shot is directly associated to an object, so that in final the shot selection ends up in selecting the objects to render, which might be difficult and irrelevant in contexts that are more complex than a video conference or a multiplayer game tv show.

Compared to all above scenarios, such as meeting room and athletic video, we target at people-dense, high activity team-sports, which not only increases the difficulty in automatic detection of salient objects but also requires a more complicated mechanism for viewpoint/camera selection. Our approach of autonomous video production is different from these previous methods in the following three aspects: (1) We perform viewpoint selection along with camera selection. (2) In our methods, we do not perform view switching between several key objects, but select the proper viewpoint based on the joint processing of the positions of multiple detected objects. (3) We care about the visual comfortness of produced sequences as well as accurate selection in each individual frame, to best render the action occurring in the covered area according to user preferences and trade-off between closeness, completeness and smoothness that we have defined in this paper.

5. Conclusion

An autonomous system for producing personalized videos from multiple camera views has been proposed. We discussed automatic adaptation of viewpoints with respect to display resolution and scenario contents, data fusion within multiple camera views, and smoothness of viewpoint and camera sequences for fluent story-telling. By performing subjective tests, we have verified the relevance of our key concepts and demonstrated the effectiveness of our approach. Furthermore, subjective tests have revealed that production of personalized video is desired by various users to meet their different viewing requirements, e.g, the different expectation of coaches and general public to the produced videos, which is a strong motivation to make video production automatic.

There are four major advantages of our methods: (1) Semantic oriented. Rather than using low-features such as edges or appearance of frames, our production is based on semantic understanding of the scenario, which could deal with more complex semantic user preference. (2) Computationally efficient. We take a divide-and-conquer strategy and consider a hierarchical processing, which is efficient in dealing with long video contents because its overall time is almost linearly proportional to the number of events included. (3) Genericity. Since our sub-methods in each individual steps are all independent from the definition of salient objects and interests, this framework is not limited to basketball videos, but able to be applied to other controlled scenarios. (4) Unsupervised. Although there are some parameters left to set by users, the system is unsupervised.

The present work also leaves the space for further improvements. We will search for an improved criterion for analytically solvable selection of viewpoints, which allows better representation for salient objects such as using motion particles or flexible body models instead of simple bounding boxes, with a manageable computation burden. Furthermore, we separate the selection and smoothing of viewpoints and cameras into four sub-steps in the current version to simplify the formulation. However, they should be solved in a unified estimation because their results affect each other. We also need to find more supports for our selection criteria

of viewpoint and cameras and search for optimal parameters by performing deeper subjective evaluations.

Acknowledgments

This work has been partly funded by FP7 Apidis European Project and Belgium NSF.

References

[1] 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>.
[2] S. Yaguchi, H. Saito, Arbitrary viewpoint video synthesis from multiple uncalibrated cameras, IEEE Trans. Syst. Man. Cybern. B 34 (2004) 430–439.
[3] N. Inamoto, H. Saito, Free viewpoint video synthesis and presentation from multiple sporting videos, Electron. Commun. Jpn. (Part III: Fundam. Electron. Sci.) 90 (2006) 40–49.
[4] I.H. Chen, S.J. Wang, An efficient approach for the calibration of multiple PTZ cameras, IEEE Trans. Automat. Sci. Eng. 4 (2007) 286–293.
[5] P. Eisert, E. Steinbach, B. Girod, Automatic reconstruction of stationary 3-D objects from multiple uncalibrated camera views, IEEE Trans. Circ. Syst. Video Technol. 10 (1999) 261–277 (Special issue on 3D video technology).
[6] A. Tyagi, G. Potamianos, J.W. Davis, S.M. Chu, Fusion of Multiple Camera Views for Kernel-Based 3D Tracking, WMVC'07, February 2007, Austin, Texas, USA, 2007, p. 1.
[7] B. Suh, H. Ling, B.B. Bederson, D.W. Jacobs, Automatic Thumbnail Cropping and its Effectiveness, ACM UIST'03, November 2003, Vancouver, Canada, 2003, pp. 95–104.
[8] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259.
[9] X. Xie, H. Liu, W.Y. Ma, H.J. Zhang, Browsing large pictures under limited display sizes, IEEE Trans. Multimedia 8 (2006) 707–715.
[10] Y. Ariki, S. Kubota, M. Kumano, Automatic Production System of Soccor Sports Video by Digital Camera Work Based on Situation Recognition, ISM'06, December 2006, San Diego, USA, 2006, pp. 851–860.
[11] J. Owens, Television Sports Production, fourth ed., Focal Press, 2007.
[12] D. Lee, Effective Gaussian mixture learning for video background subtraction, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 827C832.
[13] N. Ohta, A Statistical Approach to Backgound Suppression for Surveillance Systems, ICCV'01, July 2001, Vancouver, Canada, 2001, pp. 481–486.
[14] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, H. Wechsler, Tracking groups of people, Comput. Vis. Image Understand. 80 (2000) 42–56.
[15] I. Haritaoglu, D. Harwood, L.S. Davis, W4:real-time surveillance of people and their activities, IEEE Trans. Pattern Anal. Mac. Intell. 22 (2000) 809–830.
[16] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, Detecting moving objects, ghosts and shadows in video stream, IEEE Trans. Pattern Anal. Mac. Intell. 25 (2003) 1337–1342.
[17] J.C. Nascimento, J.S. Marques, Performance evaluation of object detection algorithms for video surveillance, IEEE Trans. Multimedia 8 (2006) 761–774.
[18] D. Delanny, N. Danhier, C. De Vleeschouwer, Detection and Recognition of Sports (wo)man From Multiple Views, ICDSC'09, August 2009, Como, Italy.
[19] A. Alahi, Y. Boursier, L. Jacques, P. Vandergheynst, Sport Players Detection and Tracking With a Mixed Network of Planar and Omnidirectional Cameras, ICDSC'09, August 2009, Como, Italy.
[20] M. Taj, A. Cavallaro, Multi-Camera Track-Before-Detect, ICDSC'09, August 2009, Como, Italy.
[21] J.W. Gibbs, Elementary Principles in Statistical Mechanics, Ox Bow Press, 1981.
[22] D. Chandler, Introduction to Modern Statistical Mechanics, Oxford University Press, 1987.
[23] C. De Vleeschouwer, F. Chen, D. Delannay, C. Parisot, C. Chaudy, E. Martrou, A. Cavallaro, Distributed Video Acquisition and Annotation for Sport-Event Summarization, NEM Summit'08, October 2008, Saint-Malo, France.
[24] M. Al-Hames, B. Hornler, R. Muller, J. Schenk, G. Rigoll, Automatic Multi-Modal Meeting Camera Selection for Video-Conferences and Meeting Browsers, ICME'07, July 2007, Beijing, China, 2007, pp. 2074–2077.
[25] R. Kubicek, P. Zak P. Zemcik, A. Herout, Automatic Video Editing for Multimodal Meetings, ICCVG'08, November 2008, Warsaw, Poland, 2008, pp. 1–12.
[26] N. Papaoulakis, N. Doulamis, C. Patrikakis, J. Soldatos, A. Pnevmatikakis, E. Protonotarios, Real-Time Video Analysis and Personalized Media Streaming Environments for Large Scale Athletic Events, AREA' 08, October 2008, Vancouver, Canada, 2008, pp. 105–112.
[27] Y. Rui, A. Gupta, J.J. Cadiz, Viewing Meetings Captured by an Omni-Directional Camera, ACM CHI'2001, March 2001, Seattle, USA, 2001, pp. 450–457.
[28] D. Vronay, S. Wang, D. Zhang, W. Zhang, Automatic Video Editing for Real-Time Multi-Point Video Conferencing, US Patent 20060251384, 2006.
[29] D. Vronay, S. Wang, D. Zhang, W. Zhang, Automatic Video Editing for Real-Time Generation of Multiplayer Game Show Videos, US Patent 20060251383, 2006