

# AUTONOMOUS PRODUCTION OF BASKETBALL VIDEOS FROM MULTI-SENSORED DATA WITH PERSONALIZED VIEWPOINTS

Fan CHEN and Christophe De Vleeschouwer

Laboratoire de Télécommunications et Télétection  
Ecole Polytechnique de Louvain, Université catholique de Louvain

## ABSTRACT

We propose an autonomous system for personalized production of basketball videos from multi-sensored data under limited display resolution. Especially, we propose criteria for optimal planning of viewpoint coverage and camera selection for improved story-telling and perceptual comfort. By using statistical inference, we design and implement the estimation process. Experiments are made to verify the system, which shows that our method efficiently alleviates flickering visual artifacts due to viewpoint switching, and discontinuous story-telling artifacts.

## 1. INTRODUCTION

We propose a computationally efficient system for producing personalized sport videos in the divide-and-conquer paradigm. By considering general production principles of sports video[1], 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 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.

In contrast to previous methods, such as threshold based optimal cropping region detection in Ref.[2], planning of optimal shifting path in Ref.[3], and soccer sport video generation in Ref.[4], our method has several advantages: it deals with the multi-camera environment; it enables to select the viewpoint adaptively as a function of user preferences, e.g., display resolution or preferred cameras; and it considers perceptual comfort as well as efficient integration of contextual information.

In Section 2, we explain the estimation framework of both selection and smoothing of viewpoints and camera views, and briefly introduce their formulation and implementation. In Section 3, experiments are made to verify the efficiency of

our system. Finally, we conclude this work and explore some paths for future research.

## 2. AUTONOMOUS PRODUCTION OF PERSONALIZED BASKETBALL VIDEOS FROM MULTI-SENSORED DATA

Since we usually locate dramatic viewpoint or camera switching during the gap between two highlighted events[1], we 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 followed, 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. Especially, we divide a game into a sequence of non-overlapped ball-possession periods according to whether the same team holds the ball, and consider the period-level continuity of viewpoint movement.

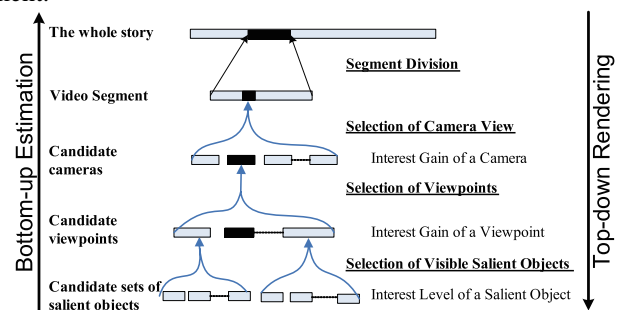


Fig. 1. Hierarchical working flow of personalized production.

Input data fed into our system include video data, associ-

ated meta-data on salient objects, 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  $i$ -th frame captured at time  $t_i$ ,  $M_i$  different salient objects, such as players, ball, referee and others, denoted by  $\{\mathbf{o}_{im} | m = 1, \dots, M_i\}$ , are detected (e.g., based on conventional video analysis tools like background subtraction) in total from all camera views. We define the  $m$ -th salient object as  $\mathbf{o}_{im} = [\mathbf{o}_{kim} | k = 1 \dots K]$ , where  $\mathbf{o}_{kim}$  is the  $m$ -th salient object in the  $k$ -th 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  $k$ -th camera view, we set  $\mathbf{o}_{kim}$  to the empty set  $\phi$ . With  $\mathbf{r}_1$  and  $\mathbf{r}_2$  being two arbitrary regions, we define several elemental functions:

$$\text{Area} : \mathcal{A}(\mathbf{r}_1) = \sum_{\mathbf{x} \in \mathbf{r}_1} 1; \quad (1)$$

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

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

$$\text{Distance} : \mathcal{D}(\mathbf{r}_1, \mathbf{r}_2) = \|\mathcal{C}(\mathbf{r}_1) - \mathcal{C}(\mathbf{r}_2)\|. \quad (4)$$

We also define user preference by a parameter set  $\mathbf{u}$ .

## 2.1. Selection of Camera/Viewpoint on Individual Frames

Some good practice principles about sport event production drives us to define a criterion for selecting an optimal viewpoint. 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. We let the viewpoint for scene construction in the  $i$ -th frame of the  $k$ -th camera be  $\mathbf{v}_{ki}$ , which is a rectangular region. For each  $\mathbf{v}_{ki}$ , we have only three free parameters, i.e., the horizontal center  $v_{kix}$ , the horizontal center  $v_{kiy}$ , and the width  $v_{kiw}$ , to tune if we fix the aspect ratio. Individual optimal viewpoint is obtained by maximizing the interest gain of applying viewpoint  $\mathbf{v}_{ki}$  to the  $i$ -th frame of the  $k$ -th camera, which is defined as a weighted sum of attentional interests from all visible salient objects in that frame, i.e.,

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

where  $\mathcal{I}(\mathbf{o}_{kim} | \mathbf{u})$  is the pre-assigned interest of a salient object  $\mathbf{o}_{kim}$  under user preference  $\mathbf{u}$ .

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,

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

where we use  $u^{\text{DEV}}$  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  $\mathcal{V}(\mathbf{o}_{kim} | \mathbf{v}_{ki})$  which separates positive interests from negative interests at the border of viewpoint; and the appended fraction part  $\ln \mathcal{A}(\mathbf{v}_{ki})$  which calculates the density of interests to evaluate the closeness and is set as a logarithm function. We let  $\hat{\mathbf{v}}_{ki}$  be the optimal viewpoint computed individually for each frame, i.e.,

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

Some optimized  $\hat{\mathbf{v}}_{ki}$  under different resolution are in Fig.2.



**Fig. 2.** Selected viewpoints under different display sizes.

### 2.1.1. Selection of Camera Views for a Given Frame

We define  $\mathbf{c} = \{c_i\}$  as a camera sequence, where  $c_i$  denotes the camera index for the  $i$ -th frame. The interest gain of choosing the  $k$ -th camera for the  $i$ -th frame is evaluated by  $\mathcal{I}_i(k | \mathbf{v}_{ki}, \mathbf{u})$ , which reads,

$$\mathcal{I}_i(k | \mathbf{v}_{ki}, \mathbf{u}) = w_k(\mathbf{u}) \mathcal{R}_{ki}^{\text{CL}} \mathcal{R}_{ki}^{\text{CP}}(\mathbf{u}) \exp \left[ -\frac{(\mathcal{R}_{ki}^{\text{OC}})^2}{2} \right]. \quad (8)$$

We weights the support of current user-preference 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. Occlusion rate  $\mathcal{R}_{ki}^{\text{OC}}$  is defined the normalized ratio of the united area of salient objects with respect to the sum of their individual area, i.e.,

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

where  $\bigcup_m x_m$  calculates the union of all bounding boxes  $\{x_m\}$ . We use  $\mathcal{N}_{ki}(\mathbf{v}_{ki}) = \sum_m \mathbf{o}_{kim} \cap \mathbf{v}_{ki} \neq \phi 1$  to represent the number of visible objects inside viewpoint  $\mathbf{v}_{ki}$ . Closeness of salient objects is defined as average pixel areas used for rendering objects, i.e.,

$$\mathcal{R}_{ki}^{\text{CL}} = \log \frac{1}{\mathcal{N}_{ki}(\mathbf{v}_{ki})} \sum_m \mathcal{A}[\mathbf{o}_{kim} \cap \mathbf{v}_{ki}]. \quad (9)$$

And the completeness of this camera view is defined as the percentage of included salient objects, i.e.,

$$\mathcal{R}_{ki}^{\text{CP}}(\mathbf{u}) = \frac{1}{\sum_m \mathcal{I}(\mathbf{o}_{kim} | \mathbf{u})} \sum_{\mathbf{o}_{cim} \cap \mathbf{v}_{ci} \neq \phi} \mathcal{I}(\mathbf{o}_{kim} | \mathbf{u}). \quad (10)$$

We define the probability of taking the  $k$ -th camera for the  $i$ -th frame under  $\{\mathbf{v}_{ki}\}$  as

$$P(c_i = k | \mathbf{v}_{ki}, \mathbf{u}) \equiv \{\mathcal{I}_i(k | \mathbf{v}_{ki}, \mathbf{u})\} / \left\{ \sum_j \mathcal{I}_i(j | \mathbf{v}_{ki}, \mathbf{u}) \right\} \quad (11)$$

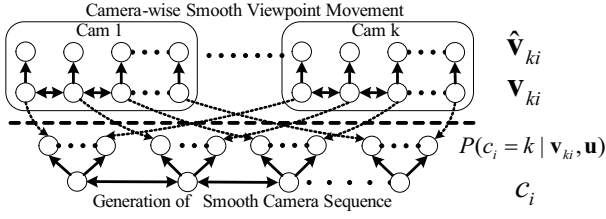


Fig. 3. Graph model for two-step viewpoint smoothing.

## 2.2. Generation of Smooth Viewpoint/Camera Sequences

A video 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.3 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.

We model both viewpoint/camera smoothing as two Markov Random Fields and using statistical physics to find the optimal configuration. We only give results. Optimized  $v_{kix}$  reads

$$v_{kix}^* = \langle v_{kix} \rangle = \frac{\sum_{j \in \mathcal{N}_i} \sigma_{2x}^2 \langle v_{kix} \rangle + \hat{v}_{kix} \sigma_{1x}^2}{\sum_{j \in \mathcal{N}_i} \sigma_{2x}^2 + \sigma_{1x}^2}. \quad (12)$$

where  $\sigma_{1x}$  and  $\sigma_{2x}$  are two parameters to control the smoothing strength and  $\mathcal{N}_i$  is the neighborhood of frame  $i$ .  $\langle x \rangle = \sum_{\{\mathbf{v}_{ki}\}} x P(\{\mathbf{v}_{ki}\} | \mathbf{u}, \{\hat{\mathbf{v}}_{ki}\})$  is the expectation value of a quantity  $x$ . We also derive correspondent updating rules in a similar way for  $v_{kiy}^*$ , and  $v_{kiw}^*$ . The smoothing process for camera sequences is performed by iterating the following fixed-point rule until reaching convergence,

$$\langle \delta_{c_i, k} \rangle^C = \frac{\exp \left\{ (1 - \gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j, k} \rangle^C + \gamma \rho_{ki} \right\}}{\sum_k \exp \left\{ (1 - \gamma) \sum_{j \in \mathcal{N}_i} \alpha_{ij} \langle \delta_{c_j, k} \rangle^C + \gamma \rho_{ki} \right\}}. \quad (13)$$

where  $\langle x \rangle^C = \sum_{\{c_i\}} x P(\{c_i\} | \{\mathbf{v}_{ki}^*\}, \mathbf{u})$  and  $\gamma$  is the smoothing strength.  $\delta_{c_i, k}$  is the Kronecker delta function.  $\rho_{ki} =$

$\log P(c_i = k | \hat{\mathbf{v}}_{ki}, \mathbf{u})$  and  $\alpha_{ij}$  normalizes the relative strength of smoothing with respect to the size of neighborhood, which reads

$$\alpha_{ij} = \frac{K}{|j - i| \sum_{l \in \mathcal{N}_i} \frac{1}{|l - i|}}. \quad (14)$$

## 3. EXPERIMENTAL RESULTS AND DISCUSSIONS

A short video clip from seven cameras with about 1200 frames is used to demonstrate behavioral characteristics of our system, especially its adaptivity under limited display resolution. In Fig.4, samples images from all the seven cameras are given. This clip covers three ball-possession periods and



Fig. 4. Sample views gathered by different cameras.

includes five events in total. In Fig.5, we show time spans of all events, whose most highlighted moments are also marked out by red solid lines. In the present paper, we evaluate our methods on manually collected meta-data for salient objects. Numerical results are depicted by graphs here while their corresponding videos are available in the website of our project [5]. Reviewers are invited to download video samples produced based on different user preferences to subjectively evaluate the efficiency and relevance of the proposed approach. Some parameters are heuristically determined based on subjective evaluation, such as the pre-assigned interest of each salient object. For viewpoint smoothing, we set all  $\beta_{ki}$  to 1 for camera-wise viewpoint smoothing in the following experiments. We also let  $\sigma_{1x} = \sigma_{1y} = \sigma_{1w} = \sigma_1$  and  $\sigma_{2x} = \sigma_{2y} = \sigma_{2w} = \sigma_2$ .

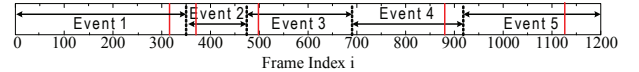
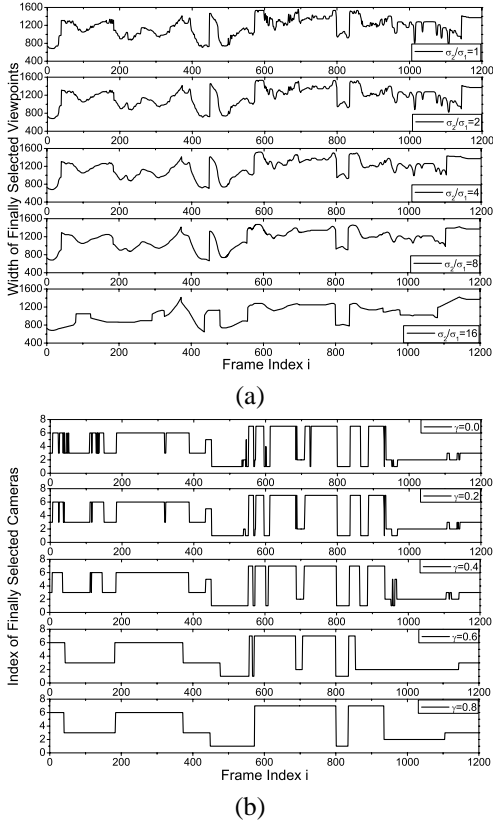


Fig. 5. A short video clip with 1200 frames is used to demonstrate the system with five clock-events inside this clip.

Viewpoint sizes of smoothed sequences under different smoothing strengths are compared in Fig.6(a). A higher ratio of  $\sigma_2$  to  $\sigma_1$  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 accented sequence which results in a flickering video. With the increasing of  $\sigma_2/\sigma_1$  ratio, the curve of viewpoint movement has less sharp peaks, whose output is perceptually more comfortable. If too strong smoothing has been performed, generated sequences will be quite different from our initial selection based on saliency information. This will cause such problems as the favorite player or the ball is out of the smoothed viewpoint. Ratio  $\sigma_2/\sigma_1$  should be determined by considering the trade-off between locally optimized viewpoints and globally smoothed viewpoint sequences.

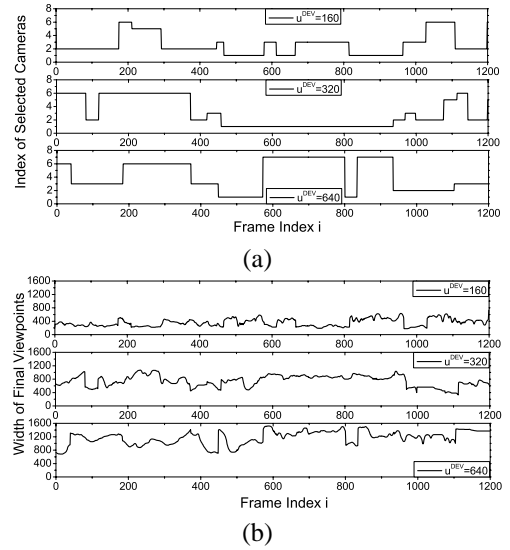
In Fig.6(b), smoothed camera sequences under various smoothing strength  $\gamma$  are depicted. A camera sequence without smoothing corresponds to the topmost sub-graph in Fig.6(b), while the sequence with the strongest smoothing is plotted in the bottom sub-graph. The unsmoothed sequence shows very annoying flickers due to dramatic camera switches, which is significantly suppressed after applying smoothing.



**Fig. 6.** Optimized camera/viewpoint sequences under different smoothing strengths with display resolution  $u^{\text{DEV}} = 640$ .

In Fig.7 (a) 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  $u^{\text{DEV}} = 160, 320$ , 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 400-th frame, the event occurs in the right court. We find that the 3-rd camera, i.e., the top-view with wide-angle lens, appears more often in the sequence of  $u^{\text{DEV}} = 640$  than that of  $u^{\text{DEV}} = 160$  and their viewpoints are also broader, which proves that a larger resolution prefers a wider view. Although the 2-nd camera appears quite often in  $u^{\text{DEV}} = 160$ , its corresponding viewpoints are much smaller in width. This camera is selected because it provides a side view of the right court

with salient objects gathered closer than other camera views due to projective geometry.



**Fig. 7.** Comparison of generated camera and viewpoint sequences under three different display resolutions 160,320 and 640, with  $\sigma_2/\sigma_1 = 4$  and  $\gamma = 0.8$ .

#### 4. CONCLUDING REMARKS

We have proposed an autonomous system for producing personalized videos from multiple camera views, which takes contextual information and outputs perceptually comfortable contents with scene coverage tailored to limited display resolution. Furthermore, our system is computationally efficient and is fully unsupervised. Currently, 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 from subjective evaluations. These will be our future works.

#### 5. REFERENCES

- [1] Owens J., "Television sports production, 4th Edition," *Focal Press*, 2007.
- [2] Suh B., Ling H., Bederson B.B., and Jacobs D.W., "Automatic thumbnail cropping and its effectiveness," *Proc. ACM UIST 2003*, pp.95-104, 2003.
- [3] Xie X., Liu H., Ma W.Y., Zhang H.J., "Browsing large pictures under limited display sizes," *IEEE Trans. Multimedia*, vol.8 pp.707-715, 2006.
- [4] Arika Y., Kubota S., and Kumano M., "Automatic production system of soccer sports video by digital camera work based on situation recognition," *ISM'06*, vol.1 pp.851-860, 2006.
- [5] Homepage of the APIDIS project and demo videos related to this paper. [http://www.apidis.org/Initial\\_Results/APIDIS%20Initial%20Results.htm](http://www.apidis.org/Initial_Results/APIDIS%20Initial%20Results.htm)