Automatic Camera Selection in the Context of Basketball Game

Florent Lefèvre[1,2(✉)], Vincent Bombardier[1], Patrick Charpentier[1], Nicolas Krommenacker[1], and Bertrand Petat[2]

[1] Université de Lorraine, CNRS, CRAN, 54000 Nancy, France
florent.lefevre@univ-lorraine.fr
[2] CitizenCam, 132 rue André Bisiaux, 54320 Maxéville, France

Abstract. This article presents an automatic video editing method for video stream selection in a multi-camera environment. The specific context of this study is Basketball game recording and broadcasting. In order to offer the best view to spectator our method is based on action detection in order to select the right camera. We use an azimuth camera to detect the center of gravity of the players representing the action in the match. The effectiveness of our method has been tested by comparing the editing obtained with that carried out by a human operator.

Keywords: Automatic editing · Detection · Sports analysis

1 Introduction

Automatic video editing allows small events to be available to a much larger audience. Indeed, many events cannot be broadcast because of the cost of the human production crew and equipment. By Automatic video editing, we mean automatic selection of the best viewing angle in a multi-camera system, in order to provide to the spectator the video stream where the action take place. CitizenCam[1] is a French company which offers multi-camera automatic recording solutions in order to retransmit on the web every type of event. Their goal is to reduce costs by automating recording and broadcasting while using IP cameras, in order to be affordable to the greatest number of people. The specific context of this study is the case of indoor sport broadcast, especially Basketball games. Automatic editing applied to sports events has been widely discussed in the literature.

Numerous methods have been developed to control virtual or real camera [5] during the recording of basketball matches [1–4,6] . Ariki et al. [1] propose an automatic production system of soccer video using virtual camera work. They track the ball in order to recognize specific events like middle shot, goal kick, corner or penalty. This event recognition allow to virtually pan and zoom on the action. However the use of virtual camera works implies not real time application.

[1] This work results from a collaboration between CitizenCam and the CRAN.

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 72–79, 2018.
https://doi.org/10.1007/978-3-319-94211-7_9

Daigo and Ozawa [6] use the fact that the audience always watches the most important part of the scene to estimate the regions of interest in a basketball match. They also use static cameras to detect the motion of the players on the court. Finally, they use these two information to generate a video by virtual panning in a panoramic video. While this method works in real time, it requires the use of a camera to capture the direction of the audience's gaze, introducing additional cost.

Chen et al. [3] propose a framework to produce personalized team-sport video summaries. They use player recognition, ball detection and event recognition in order to produce summaries, based on clock-event-based game segmentation. It's then possible to propose the spectator to select one player of interest, which will be tracked throughout the action. Although it offers a completely automatic framework, it is not possible to offer an automatic assembly in real time.

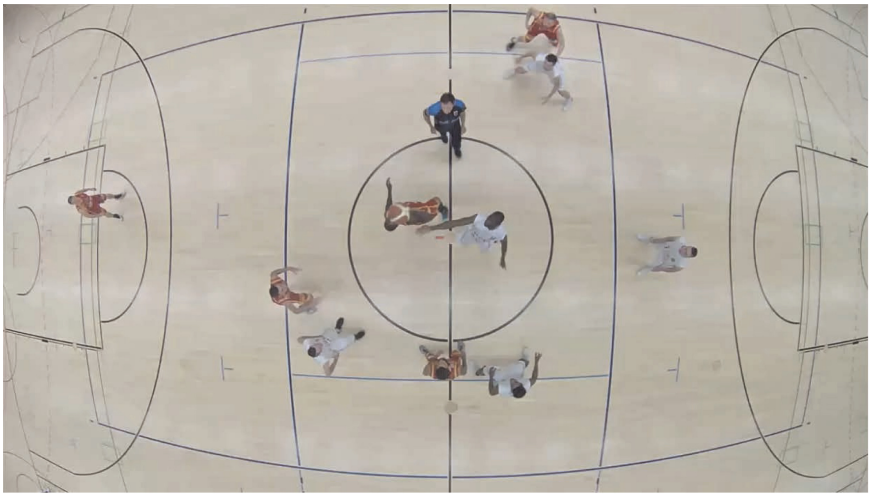
The most related works to our proposition is the one from Ren et al. [10], where they use a 8-camera system in the context of soccer games. They use images from the eight cameras, positioned at suitable locations around the stadium, to detect and track the players and the ball. The use of a large network camera reduces occlusions and allows a more efficient tracking of the ball and players. However, the scope of their paper is limited to extracting the position of the players and does not deal with the selection of a camera of interest and the use of 8 cameras is not compatible with our targeted cost.

We offer a simple method of locating the action for automatic selection of the view of interest in real time. Our system consists of only four cameras: an azimuth camera (see Fig. 1a) filming the whole field and 3 lateral cameras (Fig. 1b) filming the action at ground level. In order to offer a pleasant broadcast for the maximum number of people, we have decided to select the camera presenting the current action. One way to track the action is to follow the ball. However the ball is regularly hidden by the players' bodies (occlusions), which does not allow its continuous tracking. However, it is common that the ball is surrounded by many players, so the location of the action can also be approximated by the location of the players' center of gravity.

The first part of this paper explains the methodology used for action localization and for the camera selection. Our second step will be presenting the videos used to validate our method and the results we obtain. Finally, we introduce how this method can be improved.

2 Method Description

Since we have fixed IP camera on the side of the basketball court, the problem of automatic camera editing is reduced to a camera selection problem [5]. We use the azimuth camera in order to localize the action. Indeed, this view allows us to have the detail necessary to find the action in a match. In addition, analyzing only one video stream is less expensive than analyzing the 3 lateral video streams.



(a) Azimuth view

(b) Lateral cameras

Fig. 1. Cameras set up

The principle of our method is shown in Fig. 2 where each process is executed sequentially. In order to use this method in every situation, we introduce an initialization step to define the basketball court zones that each camera is targeting.

In order to follow the action, we decided not to use the ball as a reference, but the players' center of gravity. Indeed, we have assumed that the players' overall position, i.e. the equivalent of a center of gravity of the bounding rectangles, is representative of the position of the action. The movement of this center of gravity therefore allows you to select the side camera filming the action.

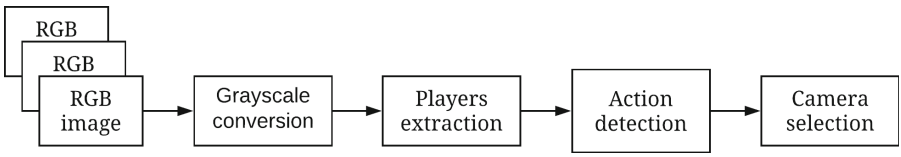


Fig. 2. Method description

2.1 Players' Extraction

Since our application focuses more on the position of the players than on their identification, color information is not useful. That's why we use grayscale images to extract players' positions. The localization of players begins with the subtraction of the background, by subtracting frames. We calculate the absolute value of the subtraction of pixels from the current image with those of the image at time t-4. Using the t-4 frame allows for accurate detection of player movement even at low player speed. Because matches are played indoors, there is not much noise disturbing the extraction of positions.

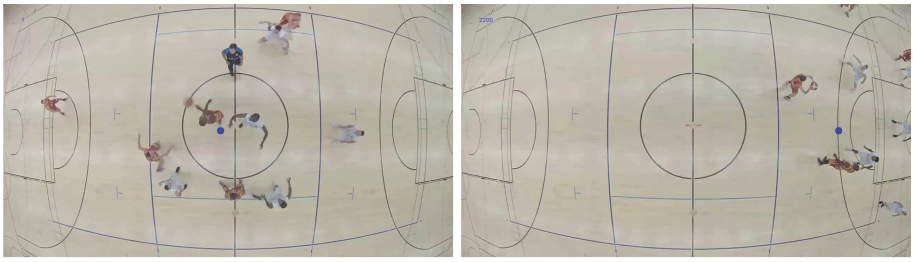
In addition, frame subtraction is efficient, in terms of calculation time than a background subtraction method (despite a decrease in accuracy): 136 fps in our case against 43 fps using a Gaussian Mixture-based Background/Foreground Segmentation [9] or 58 fps using [11].

Since artifacts may appear during subtraction, we apply a connected component analysis [8] in order to execute a dimensional thresholding. The detections which are inferior (like reflections) than a player's size are ignored. In other words, we keep only objects with a rectangle area of 400 pixels or more. This area size is the one that maximizes recall and detection accuracy.

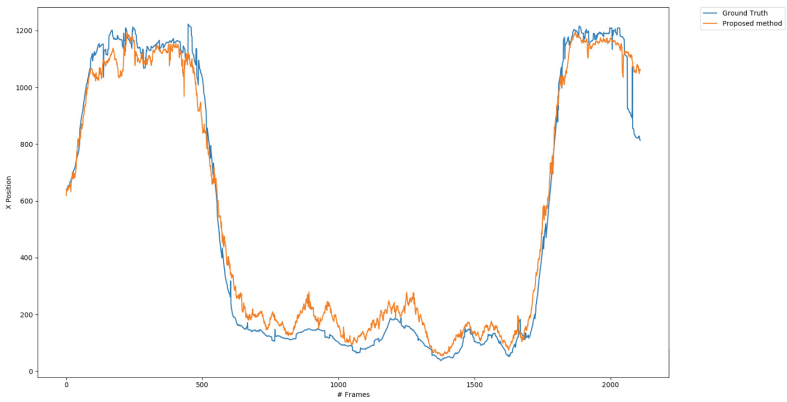
2.2 Action Detection

After analyzing the displacement center of gravity of the ground-truth, we obtain confirmation that its evolution over time could provide useful information for the selection of the camera. In the case of a basketball match, almost all the players move at the same time as the ball (see Fig. 3a). The position of the players is therefore representative of the action. It is possible that some players may stay behind while waiting for the action to return. If so, their moves will be slow, so the subtraction of the frame will not take into account these players.

The players' center of gravity corresponds to the average of each player's positions, weighted by their surface. Indeed, when several players are close (for example at the level of the basket), it is possible that some players are not correctly detected, however the detection surface increases, symbolizing the presence of several players. A comparison of the evolution of the center of gravity of our method with the ground-truth can be seen in Fig. 3b. The ground-truth has been obtained by calculating the center of gravity of all players, whose position has been manually annotated. We can notice that the center of gravity obtained by our method is close to that of the ground truth.



(a) Gravity centers in blue for the frame 1(left) and 1950 (right)



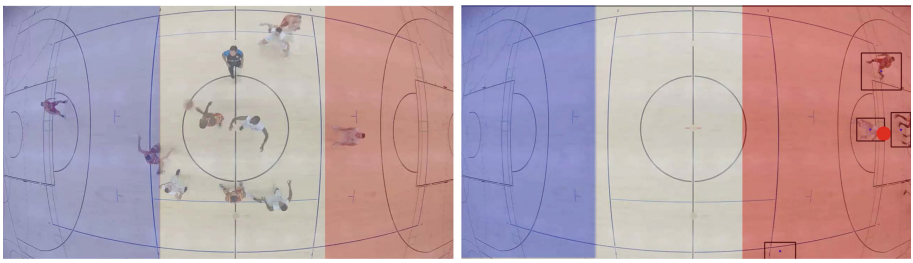
(b) Evolution of the center of gravity compared to the ground truth

Fig. 3. Evolution of the center of gravity

2.3 Camera Selection

Since we have 3 cameras, located at the edge of the field, we defined three zones in the azimuth image, as we can see in Fig. 4 on the right. As soon as the players move from one area to another, we switch cameras.

In order to avoid a lot of camera changes, as when the action happens at the basket level, we use a hysteresis function (Fig. 4 - left). Indeed, when a team tries to score a basket, it is often the case that the players move backwards in order to space the game. This rearward movement can lead to a camera change, although the action takes place under the basket. The hysteresis function keeps the camera close to the basket until a large number of players have returned to the opponent's side.



(a) (b)

Fig. 4. Zones definition: (a) the three zone defined - (b) the hysteresis function to select the view

3 Validation

The evaluation of the proposed method was performed on a computer equipped with an Intel Core I7-5557U processor (Base Frequency 3.1 GHz) and 8 GB RAM. The method was implemented in Python, using functions from OpenCV library.[2]

3.1 Video Database

The videos we used come from the match between BC Mess and Heffingen in October 2016, available on citizencam.tv. Three different sequences were used to evaluate the sequence. The first two sequences (2110 frames and 3000 frames) were used as ground-truth to validate the estimation of the center of gravity displacement. These sequences have to be manually annotated, which explains the low number of frames in these sequences. The last sequence was used in order to compare the editing obtained by our method, with a human operator editing.

3.2 Results

The processing time per image is about 7–8 ms (130 fps), which is compatible with live broadcasting. As we can see in Fig. 3b, the center of gravity of our method is close to that of the ground-truth. This confirms that we can use the center of gravity as a lever to select the right camera. Figure 5 compares the assembly obtained by our method with the one obtained using the center of gravity of the ground-truth and a manual assembly. The results show the similarity between these three fixtures. The difference with manual editing is that humans tend to anticipate the action or have a delay in selecting the camera, which is not reproducible algorithmically.

In the case of the largest sequence, 23 camera changes were performed by humans versus 21 by our method. This difference is due to the fact that the person doing the editing will occasionally look for a larger view of the action

[2] opencv.org.

when players space the game at the basket level. The individual follow-up of each player will allow to visualize this spacing in relation to the center of gravity and to propose the adequate view.

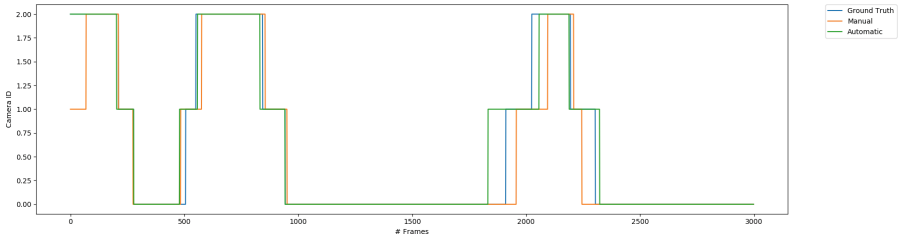


Fig. 5. Comparison of editing

4 Conclusion

We used the information from a wide-angle camera above the stage to determine which of the 3 basketball court views contains the action. Tracking gravity centers allows efficient selection of the interesting view. The results show that we obtain video editing close to those obtained with human operator. The advantage of our method is the possibility of being used in real time, which implies the possibility of being used for live broadcasting. Moreover, the ease of initialization is part of CitizenCam's desire to make event broadcasting simple.

Nevertheless, there are many improvements to be made, such as the detection of events (free-throws, lost balls, field goals,...) allowing to improve the shooting during specific actions. The monitoring of each player [7] will improve live shooting. It will be possible, for example, to identify counter attacks in order to select the camera where the attacker is, instead of waiting for all players, and therefore the center of gravity, to move. The knowledge of the position and the identification of each player, as well as the recognition of the events will also make it possible to propose, after the diffusion, an assembly or a summary specific to the preferences of each spectators.

References

1. Ariki, Y., Kubota, S., Kumano, M.: Automatic production system of soccer sports video by digital camera work based on situation recognition, pp. 851–860, December 2006
2. Carr, P., Mistry, M., Matthews, I.: Hybrid robotic/virtual pan-tilt-zom cameras for autonomous event recording. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pp. 193–202. ACM, New York (2013). https://doi.org/10.1145/2502081.2502086
3. Chen, F., Delannay, D., Vleeschouwer, C.D.: An autonomous framework to produce and distribute personalized team-sport video summaries: a basketball case study. IEEE Trans. Multimed. **13**(6), 1381–1394 (2011)

4. Chen, J., Carr, P.: Mimicking human camera operators. In: 2015 IEEE Winter Conference on Applications of Computer Vision, pp. 215–222, January 2015
5. Chen, J., Carr, P.: Autonomous camera systems: a survey. In: Workshop on Intelligent Cinematography and Editing, pp. 18–22 (2014)
6. Daigo, S., Ozawa, S.: Automatic pan control system for broadcasting ball games based on audience's face direction, p. 444. ACM Press (2004). http://portal.acm.org/citation.cfm?doid=1027527.1027634
7. Delannay, D., Danhier, N., Vleeschouwer, C.D.: Detection and recognition of sports (wo)men from multiple views. In: 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), pp. 1–7, August 2009
8. Fiorio, C., Gustedt, J.: Two linear time union-find strategies for image processing. Theoret. Comput. Sci. **154**(2), 165–181 (1996). http://www.sciencedirect.com/science/article/pii/0304397594002622
9. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino, P., Jones, G.A., Paragios, N., Regazzoni, C.S. (eds.) Video-Based Surveillance Systems, pp. 135–144. Springer, Boston (2002). https://doi.org/10.1007/978-1-4615-0913-4_11
10. Ren, J., Xu, M., Orwell, J., Jones, G.A.: Multi-camera video surveillance for real-time analysis and reconstruction of soccer games. Mach. Vis. Appl. **21**(6), 855–863 (2010). https://doi.org/10.1007/s00138-009-0212-0
11. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27** (2006). http://dare.uva.nl/search?metis.record.id=263396