Workshop on Activity Monitoring by Multiple Distributed Sensing (AMMDS) in conjunction with
2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

An automatic system for sports analytics in multi-camera tennis videos

Rafael Martín Nieto
Universidad Autónoma de Madrid
Madrid, Spain
Rafael.MartinN@uam.es

José María Martínez Sánchez
Universidad Autónoma de Madrid
Madrid, Spain
JoseM.Martinez@uam.es

Abstract

This paper presents an automatic system which after a simple previous configuration is able to detect and track each one of the players on the court or field in single player sports. After that, the system is able to extract statistics and performance of the players. This system is complete, general and modular, to be improved and modified by future work. The system is based on a mono-camera detection and tracking system, originally designed for video surveillance, which has been adapted for its use in the individual sports domain. Target sports of the developed system are individual sports (e.g., tennis, paddle tennis) where the players have its own side of the field.

1. Introduction

A lot of work has already been carried out on content-based analysis of sports videos, especially in soccer, basket and tennis, and the work on enhancement and enrichment of sports video is growing quickly due to the great demands of customers.

Since tennis is one of the most popular sports, tennis video analysis is an area with many possibilities and interesting for audiences. In sports coaching, technology can give competitive advantages to the players.

An overview of sports video research is given in [1], describing both basic algorithmic techniques and applications. Sports video research can be classified into the following two main goals: indexing and retrieval systems (based on high-level semantic queries) and augmented reality presentation (to present additional information and to provide new viewing experience to the users).

Sports video can be classified in

- Edited for broadcast [2][3][4][5]: in this type of videos, there are scenes of the game, repetitions and changes of viewpoint of the observer. This type of videos is the most commonly available.
- Non edited mono- [6][7][8] or multi-camera [9]: these video can be shot by mobile cameras, but generally they are shot by fixed cameras, as this characteristic greatly facilitates the processing.

In the case of individual sports there are multiple possibilities when analyzing the videos: player detection and tracking [8], tennis strokes detection [6] or ball detection and tracking [5][7].

For team sports the problem is more complicated because there are multiple players, usually with similar appearance. In this case, occlusions occur and must be considered. Some applications of processing these kinds of sports are: virtual view synthesis between recorded views[10], detection of the ball-player interactions[11] or evaluation of players performance and skills[12].

The presented system is centered on individual sports. With individual sports we mean sports where every player has his own area of the field and no other player or referee can enter in the area of the monitored player. This is a fundamental aspect because it allows using simple approaches for combining the views from the different cameras.

The main objective of this paper is to present an automatic system able to detect and track the players in multi-camera sports videos fusing tracks across cameras. This work is based on a mono-camera detection and tracking system designed to work for video surveillance applications and that has been adapted to operate within the sports domain.

The remainder of the paper is structured as follows: after the introduction in the first section, section 2 describes the designed system; section 3 presents the adjustments, testing and results of the implementation of the system, including the description of the used dataset; finally, section 4 describe the conclusions and future work.

2. Video analysis system

2.1. System overview

There is an individual instance of the system for each player. The videos used from the content set are those covering only the areas of each player, not global views. That is, for a full match there will be two parallel systems.

The system block diagram is depicted in Figure 1:

The Initialization block is executed first. It generates the field representation, the backgrounds, the masks and the homography reference points.

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

The *Tracking block* (mcD&T –see section 2.3-) receives the videos and the backgrounds and masks from the different cameras which are recording the player and generates the tracking from its own perspective.

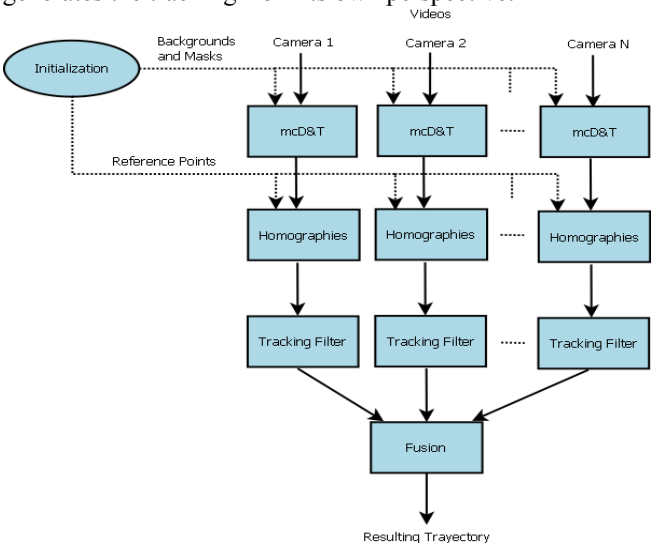

Figure 1: Block diagram of the system

The *Homographies block* receives the tracking of the player from each camera and the reference points, and it generates the top view tracking from each camera.

The *Tracking filter block* receives and filters the tracking for the top view tracking from each camera. Some examples of filtering criteria are: selected regions in the field, blob continuity for a minimum number of frames, etc. This block generates the top view filtered tracking for each camera.

Finally, the *Fusion block* joins the processed trackings from each camera and generates the overall trajectory along the entire field.

2.2. Initialization

Initialization of the system can be obtained from a previous video recorded before the match or from the first frames of the match video (if the first option is not available).

The initialization step generates the field representation, the backgrounds, the masks and the homography reference points. The measures of the field are needed for the field representation (otherwise the resulting statistics would not be accurate).

2.3. Mono-camera detection and tracking

The mono-camera detection and tracking (mcD&T) module is based on a video analysis system for event detection in the video-surveillance domain[13]. It is designed to work in real time: this requirement imposes limits on the time complexity of the algorithms used in each of the analysis modules.

Figure 2 depicts the block diagram of the mono-camera detection and tracking module. A foreground mask is generated for each incoming frame at the *Foreground Segmentation Module*. This foreground mask consists on a binary image that identifies the pixels that belong to moving or stationary blobs. Then, post-processing techniques are applied to this foreground mask in order to remove noisy artifacts and shadows. After that, the *Blob Extraction Module* determines the connected components of the foreground mask. In the following stage, the *Blob Tracking Module* associates an ID for each extracted blob across the frame sequence.

2.4. Homographies

Homographies are used to change camera perspectives. The homographies applied follows the normalized direct linear transformation algorithm[14]. Four points are selected in the original image plane (generally characteristic points of the field) and its correspondence is indicated in the top view. The code used for the homographies is public[1].In Figure 3, an example of the result is shown. Note that in the system the homography is applied to a single point (the base mid-point) of the tracked player (blob). The red points in the figure indicate the selected points.

Resulting projected trajectories from two cameras may have location differences due to different scales, players' volumes, camera distances, lens (e.g., see Figure 3), etc. In the case of people detection and tracking, these differences use to be moderate and the fusion block can also reduce

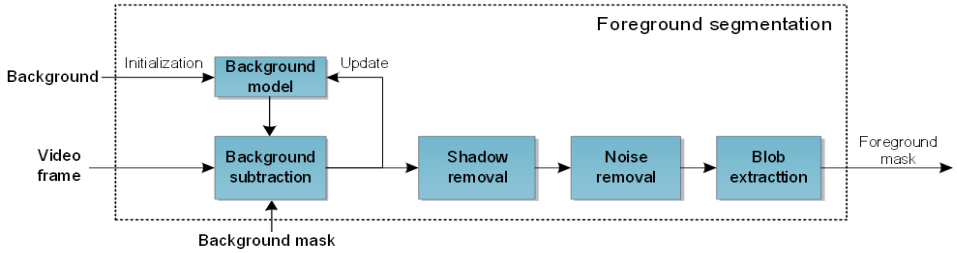

Figure 2: Block diagram of the mono-camera detection and tracking module

[1] http://www.csse.uwa.edu.au/~pk/research/matlabfns/

these errors.

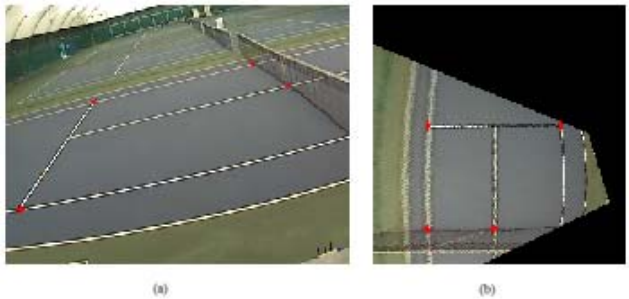

Figure 3: Example of homography: (a) image plane, (b) Top view

2.5. Fusion

For individual sports with cameras covering only individual player's field, the fusion process is relatively simple. The main advantage is that every blob in frame t belongs to the tracked player. The resulting coordinates can be obtained simple as a mean:

$$x_t = \frac{1}{N}\sum_{i=1}^{N_t} x_{i,t} \qquad y_t = \frac{1}{N}\sum_{i=1}^{N_t} y_{i,t} \qquad (1)$$

Where (x_t,y_t) are the fused coordinates in frame t, $(x_{i,t}, y_{i,t})$ are the coordinates in frame t from the i-camera, and $i=1\ldots N_t$ are the N_t cameras which have a recorded blob corresponding to the player in frame t.

Additionally, a weight can be added to each blob coordinates depending on the position, height and precision of each camera.

3. Adjustments, testings and results[2]

3.1. Content set: 3DLife ACM Multimedia Grand Challenge 2010 Dataset

The dataset[3] features video from 9 CCTV-like cameras placed at different points around the entire court (see Figure 4). In addition, audio from 7 on the 9 cameras (each camera with the microphone symbol in the top left of the image) is also available. Videos are ASF files and encoded using an MPEG-4 codec. 7 of the videos (taken from the cameras with the microphone symbol in the top left of the image) are recorded with a resolution of 640x480 pixels from Axis 212PTZ network cameras. The two other cameras have a resolution of 704x576 pixels and are captured using Axis 215PTZ network cameras. The start time of each video is synchronized via software at the start of each sequence.

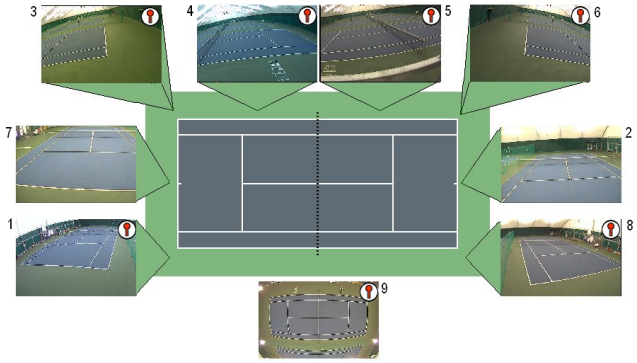

Figure 4: Positions and examples of the content set videos

3.2. Integration

For the integration, the used videos are from the cameras 1, 3 and 4 from 3DLife ACM Multimedia Grand Challenge 2010 Dataset, recording the blue player in game 1. The other videos have problems in the tracking module for many causes: the net that separates the fields is moved by the wind, camera that automatically focuses and blurs, camera slightly moved by the wind, etc. These limitations are produced by the mcD&T module. Solving these problems was not in the objectives of this work and will be described as future work.

The extractor used is a simple state of the art method (median background subtraction). The homography reference points are selected manually. In the *Tracking filter* module, the criteria (heuristically set) for filtering are that detected blobs must be visible during a minimum of 10 consecutive frames and that the detection area is limited to own field. For the videos used in the experiments, applying masks was not necessary so empty masks were generated. If necessary, masks are easily set using regions limited by points defined by the system user.

The average time of the tracking processing in the tennis videos is about 21.9 fps in a standard PC.

3.3. Results

Once the fusion module has generated the trajectory of the player, player statistics can be extracted.

A zigzag effect occurs between consecutive frames, as shown in Figure 5. As commented previously, there are multiple sources of error that cause this problem: different camera lenses, user precision in the homographies selection points, segmentation or tracking errors, etc. This effect causes an error in the obtained statistics. To reduce this error, the statistics are calculated after postprocessing the trajectory (subsampling every 25 frames and applying linear interpolation).

[2] A web page has been created to add videos and results, where videos with the obtained results of the systems have been added.
http://www-vpu.eps.uam.es/publications/AnAutomaticSystemForSportsAnalyticsInMulticameraTennisVideos/

[3] http://www.cdvp.dcu.ie/tennisireland/TennisVideos/acm_mm_3dlife_grand_challenge/

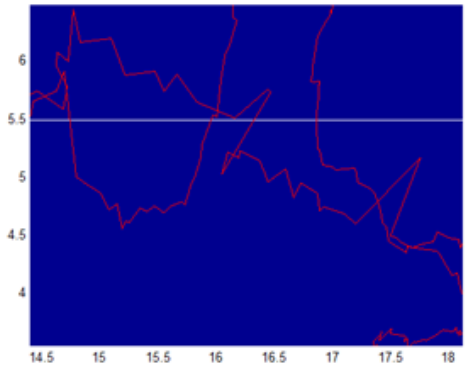

Figure 5: Example of the zigzag effect in the trajectories

An example of the resulting statistics video is shown in Figure 6, which depicts the following information: position on the field (X and Y given a coordinate origin), total covered distance, average speed and instant speed.

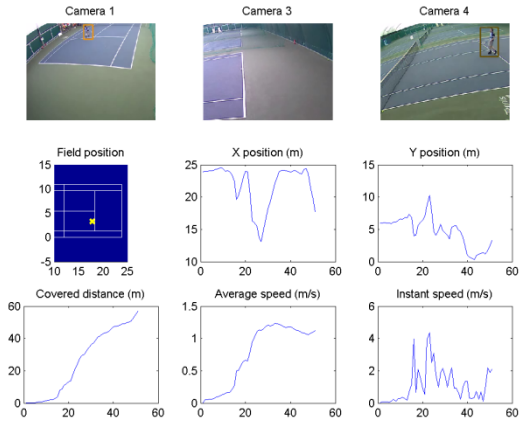

Figure 6: Example of the resulting statistic video

4. Conclusions

The main objective of the work presented in this paper, the design and development of an automatic system for detecting and tracking players in a field using multi-camera video, has been reached. After a simple configuration, the system is able to detect and track the player in each side of the field. With the resulting trajectory, the system is able to provide some statistics. The system is complete, general and modular, easing future work improvements and modifications.

The evaluation using Ground Truth is not possible since there is no such Ground Truth. The results have not been compared with other systems as we have not found software or results using the same dataset.

Analysis of sport videos with fixed cameras has advantages with respect to general videos of video surveillance. Backgrounds are generally static and uniform, except for certain areas such as public or dynamic advertising, which can be excluded easily.

The location of the cameras is important. The ideal case is when the cameras are symmetrically placed covering most of the field, since in these cases tracking errors are reduced in fusion process. Placing the cameras at higher elevations is interesting because it reduces the tracking error.

There are many lines of future work for this system. A real-time interactive Graphical User Interface (GUI) should be developed for the system operators or system supervisors. The precision of each trajectory projection depends on the location in the field. An evaluation of the precision as a function of the position in the field and of the distance to the camera can prevent these precision differences, reducing the final position error. The applied fusion is one of the simplest possible techniques. Some more complex techniques can be applied: as a first approach, adding weights to the camera projections depending on the distance to the player in each frame. Finally, additional statistics and information can be calculated: areas of the field where the player stays longer, instant acceleration of the player, etc.

Acknowledgements

This work has been partially supported by the Spanish Government (TEC2011-25995).

References

[1] Y. Xinguo, D. Farin. Current and emerging topics in sports video processing. In Proc. of ICME 2005, 526-529.

[2] Y. Huang, J. Llach, S. Bhagavathy. Players and ball detection in soccer videos based on color segmentation and shape analysis. In Proc. of ICMCAM 2007, vol. 4577: 416-425.

[3] C. Poppe, S.D. Bruyne, S. Verstockt, R.V. de Walle. Multi-camera analysis of soccer sequences. In Proc. of AVSS 2010, 26-31.

[4] M. Xu, J. Orwell, G. Jones. Tracking football players with multiple cameras. In Proc. of ICIP 2004, vol. 5:24-27.

[5] X. Yu, C.H. Sim, J.R.Wang, L.F. Cheong. A trajectory-based ball detection and tracking algorithm in broadcast tennis video, In Proc. of ICIP 2004, vol. 2:1049-1052.

[6] D. Connaghan, C. Ó Conaire, P. Kelly, and, N.E. O'Connor. *Recognition of tennis strokes using key postures.* In Proc. of ISSC 2010, 23-24.

[7] F. Yan, W. Christmas and J. Kittler. A tennis ball tracking algorithm for automatic annotation of tennis match. In Proc. of BMVC 2005, 619-628.

[8] Y.C. Jiang, K.T. Lai, C.H. Hsien, M.F. Lai. Player Detection and Tracking in Broadcast Tennis Video. In Proc. of PSIVT 2009, 759-770.

[9] G. Kayumbi, P.L. Mazzeo, P. Spagnolo, M. Taj, A. Cavallaro. Distributed visual sensing for virtual top-view trajectory generation in football videos. In Proc. of CIVR 2008, 535-542.

[10] N. Inamoto, H. Saito. Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. IEEE Transactions on Multimedia, 9(6):1155–1166, 2007.

[11] M. Leo, N. Mosca, P. Spagnolo, P. L. Mazzeo, T. D'Orazio, A. Distante. A Visual Framework for Interaction Detection in Soccer Matches. In International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 24(4):499-530, 2010.
[12] P.S. Tsai, T. Meijome, P.G. Austin. Scout: a game speed analysis and tracking system. Machine vision and Application, 18(5):289–299, 2007
[13] Juan C. SanMiguel, José M. Martínez. A semantic-based probabilistic approach for real-time video event recognition, Computer Vision and Image Understanding, 116(9): 937-952, 2012.
[14] R. Hartley , A. Zisserman. A Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.