A Computer Vision-based Information Retrieval System for Soccer Videos

Quang Tran, Tien Dinh, Duc Duong
Faculty of Information Technology
University of Science, Vietnam National University – Ho Chi Minh
Ho Chi Minh city, Vietnam
{tmquang, dbtien, daduc}@fit.hcmus.edu.vn

Abstract—**The paper proposes an information retrieval system for broadcast soccer videos using computer vision techniques. The system has been divided into 3 main phases. Firstly, the input videos have been analyzed to remove noises caused by the audience. Only the playfield and players are extracted. The system detects the players based on their appearance features. Secondly, the system tracks all the players in the video frames by using a multi-target tracker. Concurrently, the field model is built on a fast calibration algorithm by fitting two regions: the center circle and the penalty areas so as to provide the geometry transformation called homography that enables to map each player position in the video frame to the real-world coordinate. Finally, the system runs an optical character recognition module to capture and extract the text information provided in the broadcast soccer videos, such as the current scores, the name of players in a substitution or in a foul event, and the name of the scorer. The 3-phase computer vision-based system will extract a lot of meaningful information from the games. The presented framework is evaluated on the 2010 FIFA World Cup South Africa with various environments and challenging conditions, as well as compared with some other previous work in many aspects.**

Keywords – detection and tracking; soccer videos; information retrieval

I. INTRODUCTION

With the explosive growth of digital video data and computing power, sports events especially become the most popular media in many parts of world. Consequently, the software applications such as storing, arranging and retrieving of the video data are required. To be quite exact, the basis of the above computer program is a sports analysis framework. In this literature review, we would like to concentrate on human detection, tracking algorithms and related applications in video analysis.

A. Human Detection and Tracking

Among sports research domains, human detection and tracking present two fundamental steps in many vision systems. Early works of human detection [1][2][3][4] are based on background subtraction techniques, that rely on the information of background model and then perform the comparison of each frame by thresholding to gain the foreground blobs. To identify human body, the structure of blobs is examined. This idea has own benefits – fast operation and less complexity but its drawback is that it is hard to build a solid background in case of camera moving. Some techniques also claim to build adaptive background solving the moving screen such as running average [1], however, they also get stuck to detect people who have stood on one place for a long time, i.e. they are considered as background models. Recently, some efficient approaches provide a better result on detecting pedestrian. Viola et al. [5] proposed an integral image technique and Adaboost algorithm that would be done in extracting and learning features. Whilst, Dalal and Triggs [6] proposed a set of features namely the Histogram of Oriented Gradients that perform well in detecting people by using a linear support vector machine classifier. Pictorial structure models [7] have been used for detection of body parts and for articulated object-based recognition.

In addition, tracking approach becomes powerful in exploiting the trajectory of objects. Some probabilistic tracking methods can be used popularly, especially, Particle filter [8][9] approaches have been developed and improved from the original Condensation [10]. To point out the state of the art tracking, Zdenek et al. [11] devised a novel tracking-learning-detection framework, which uses a detector based on random forest to scan whole image and the best one decided by the online object model is the current target. To achieve to track multiple humans, single object tracking could not be applied well. Many multiple tracking models [12][13][14] have been proposed to deal with multiple targets mainly supporting for visual surveillance, and not been applied to sports.

B. Video Analysis for Sports

Extensive research efforts recently have contributed to many types of sports. Farin et al. [15] presented a robust camera calibration for semantic analysis of sports to provide the geometry transformation between image and real-world coordinate. Inspired from this calibration technique, Dang and Tran [17] provided a real time player tracking system for tennis broadcast that uses a heuristic approach to enhance the tracking process by considering the distracters – medium size foreground blobs. In terms of soccer, Yu et al. [18][19] proposed a trajectory-based algorithm for detecting and tracking the ball in soccer video in order to analysis play-break structure, team ball possession as well as detect basic actions. Detection and tracking of objects are much more difficult mainly due to unfixed cameras and various directions of object

movement. Some modern technologies [5][6] could be applied, but they require some particular conditions like time training for [5] and high resolution images for [6]. In brief, we believe that [20] would be a flexible method for player segmentation because it is widely applied in many kinds of image sizes and has less time training.

C. The proposed approach

Our system consists of two main processes – Offline; Query and Visualization shown in figure 1. When a specific soccer tournament was selected (in this paper, we used the 2010 FiFa World Cup dataset), as a matter of fact, the configuration of all broadcast videos is identical (frame size, on-screen text and video quality), we then build the storage file by running offline via those components – segmentation and tracking, field model, optical character recognition (OCR). From an XML file and its tournament dataset, an user could retrieve valuable information like shot views, player location on long-vew shot, and events. To provide a visual image of players on real world model, a viewer could choose an option in this case showing the penalty areas or the circle center where the planar tracjectory of players is projected.

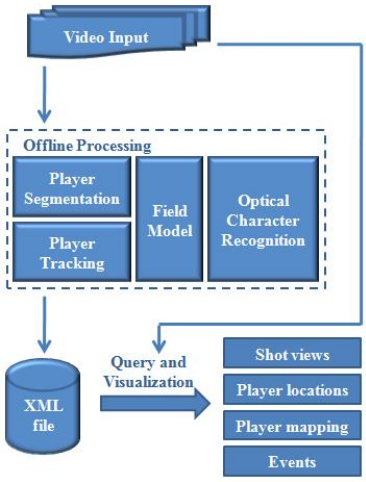

Figure 1. An overview of our framework.

The rest of the paper is organized as follows. Player segmentation and tracking are described in Section 2. Field model and Optical Character Recognition are then presented in Section 3 and Section 4 respectively. Experiment results are shown in Section 5, followed by the conclusion in Section 6.

II. PLAYER SEGMENTATION AND TRACKING

A. Player Segmentation

In this section, we introduce our segmentation inspired by the long-view soccer player detection framework [20]. A test sequence comprised of different shot views – long views, short views, out-of-field views, close-up views can be classified by rules which are based on the playfield extracting method of Ekin *et al.* [28]. For the long-view shots, the players are detected by analyzing edge features, this process sets up an initial stage for further tracking wok.

B. Player Tracking

Each players are represented as $\mathbf{x} = (x,y,h,w,id)$ where (x,y) is the position, (w,h) is rectangle size and (id) is ID number. Our implementation is entirely adopted reversible jump Markov chain Monte Carlo based method, in more details [21], applied for soccer player tracking by constructing the samples on the state space with the current state X_t and the posterior probability $p(X_t|Z_{1:t})$ where $Z_{1:t}$ is the observation state up to time t. Maximum A Posterior (MAP) is estimated:

$$\hat{X}_t^{MAP} = \underset{X_t}{\operatorname{argmax}}\, p(X_t|Z_{1:t}) \quad (1)$$

In association process, the identification number of a player is determined based on link between the current position and the previous one. If i^{th} object at time t links to j^{th} object at time $t-1$, we have $link_t(i) = j$, intuitively i^{th} object at time t and j^{th} object at time $t-1$ is one object. If i^{th} object at time t has no link with others, we have $link_t(i) = -1$ meaning a new object. Conversely, if j^{th} object at time $t-1$ does not have any link to others at time t, this means the object moves out of view.

III. FIELD MODEL

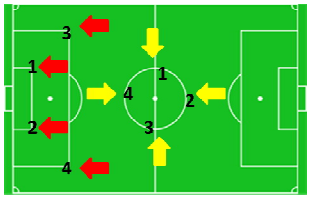

Figure 2. Field model: four penalty points (red arrows) and four center circle points (yellow arrows).

Since the field and the displayed image are planar, we use a homography that considered as a *3x3* transformation matrix **H**. An image point p = (x, y) is related to model point p' = (x', y') by **p' = Hp**.

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \mathrm{H} \times \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} \quad (2)$$

In penalty areas and center circle matching, we try to match four points in the image and four points in the real-world coordinate. With enough four correspondence points between the reference and the target images, eight coefficients (dividing each term by h_{33}) can be determined by the Gaussian elimination method. Two pairs of four points in field model are illustrated in figure 2.

A. Penalty area matching

1) White-pixel field extraction

The colour of field-lines in soccer game are always white, a threshold method based on luminance distance can be applied to extract white pixels with an additional constraint to reject amount of white areas from the detection result [15]. A field line pixel satisfies two conditions – its luminance exceed a threshold σl and the darker pixels are checked at the distance of τ pixels from surrounding areas of the candidate pixel (τ is the approximate line width). This indicates that white field-line

pixels must be enclosed either horizontally or vertically by dark pixels.

2) Line parameter estimation

Based on the set of white field-line pixels, field-line will be detected by RANSAC algorithm [22]. We collect the dominant line and remove white pixels along line segments from the dataset in an iterative process, those segments are further likely connected together to make longer segments. From the set of candidate lines, we classify them into two classes – vertical and horizontal lines.

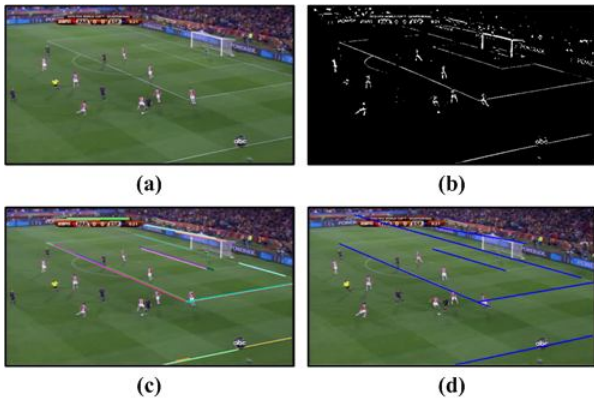

Figure 3. (a) Current frame, (b) White pixels, (c) Detected lines (each line is represented by each colour) and (d) Line candidates.

RANSAC is randomized algorithm that hypothesizes a set of model parameters and verifies the quality of the parameters. Several hypotheses have been evaluated to choose the best one. An evaluation score of the line parameters is defined as:

$$s(g) = \sum_{(x',y')\in P} \max(\tau - d(g, x', y'), 0) \quad (3)$$

Where P is the set of field-line pixels and *d(g,x',y')* is the distance of pixel *(x,y)* to line g. This high score line indicates the strong support of its hypothesis as the number of white pixels close to. The score also determine the two end points of the line.

From those detected lines, two neighbor lines are emerged to each other if their angle is smaller than 0.75^{o} and their distance is less than 5. This refinement step aims to reduce the number of possible hypotheses in calibration parameter searching space; therefore, improve the algorithm performance. A sample result is depicted in figure 3.

3) Line parameter estimation

The model fitting task determines correspondences between detected lines and the lines in the field model. We try to construct the transformation parameters for each correspondence of line segments in the image and the model. For each construction of lines, we use the proposed method of matching score [24] to find the best homography. Note that the computational cost for each calibration setting is low, but the scoring of the supporting model is very expensive; hence, some heuristic tricks to reject impossible parameters are applied.

B. Center circle searching

1) Ellipse fitting

Center circle is always white; we use the same method in section 6.1.1 to extract white-pixels. Based on the proposed method of detecting arc [23], we apply the Advance Least Square Fitting to find an initial ellipse for each contour. Then the initial ellipse will be swollen by the surrounding white pixels

2) Line parameter estimation

In several ellipses generated, an ellipse with the highest score is selected and satisfies the condition below:

$$score = \sum_{\substack{all\ pixels\ (x,y) \\ in\ ellipse}} \begin{cases} 2 & if \text{ the pixel(x,y) is a white court pixel} \\ -1 & if \text{ the pixel(x,y) is not a white pixel} \\ 0 & if \text{ the pixel(x,y) is outside the image} \end{cases} \quad (4)$$

Two rules should be checked:

$$\begin{cases} a > b > \dfrac{a}{10} \\ |\theta| < \dfrac{\pi}{4} \end{cases} \quad (5)$$

The homography matrix is computed by four extreme points of the detected ellipse in image and four given points in the model.

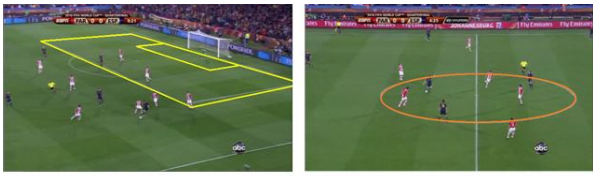

Figure 4. Matching results:(a) Penalty area, (b) Center circle.

An example is shown in figure 4. Based on the homography matrix, the trajectory of soccer player in section 2 can be mapped on the ground plane of field model. It is much more useful to support for tactic information analysis.

IV. OPTICAL CHARACTER RECOGNITION

Soccer videos often include information in text, which is often ignored by previous systems, corresponding to special events such as goals, substitutions, yellow cards and red cards. Based on these information, we can answer some questions such as who is the scorer? Who is going to be in and off the field? Who has got a yellow card or red card? To retrieve this fruitful information, we propose a simple and efficient method to detect these text captions in soccer videos by assuming that they come from the same television station.

From these videos, we choose some sample text images of each event and store its appearance location and size. We assume that the location and size of each event in the same tournament is not changed for all videos from the same broadcaster. Moreover, text caption for each event always appears in a large number of continuous frames. Based on these characteristics, we firstly use histogram matching to detect possible text caption for each frame. For each event, when the number of consecutive frames are large enough (chosen as 100), we apply a refine processing method to extract text information.

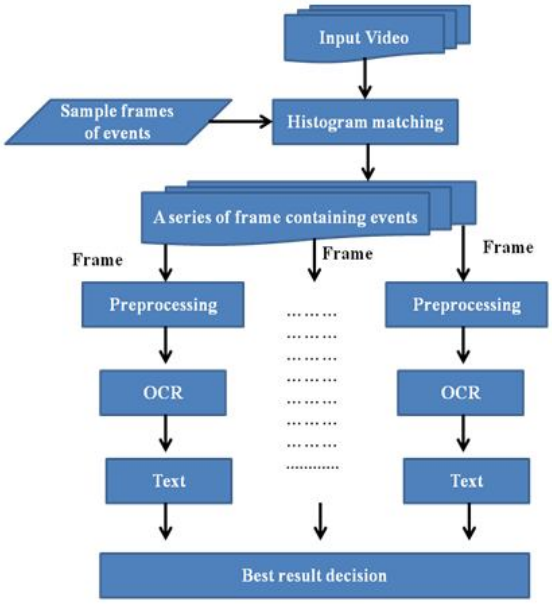

Figure 5. Flowchart of retrieving text information from soccer videos.

For each text image from a series of frames, we do the following steps, depicted in figure 5:

- Zone location information.
- Zoom in each text area from 2 to 5 times to isolate characters if the text area is too small.
- Convert scaled image to gray scale image.
- Segment scaled image into K regions using K-Means algorithm. In our system, based on experiments and suggested in [24], we choose K= 3.
- Assign black color for pixels which are closest to the letter color (usually black or white color). The output of this step is a binary image which only contains black pixels that belong to text image.
- Binary images are then processed by OCR module to get text information. After that, we make statistic text results from continuous frames of each event. The final result is the text having high appearance frequency in the continuous frames. In OCR module, we use Neuron Network to train characters and numbers. From then, it is used to recognize the text in an image.

Figure 6. OCR results: (a) Match, (b) Goals, (c) Substitution, and (d) Yellow card.

Figure 6 displays the sample result of Argentina – Mexico match. The method looks completely successful in extracting text information under significant variations.

V. EXPERIMENTAL RESULTS

In this section, We have tested our framework on video sequences recorded from the 2010 FiFa World Cup South Africa that demonstrate the performance of the proposed approach for two pronged components: tracking-mapping components and OCR module when they runs offline. In retrieval, the whole system could run real-time, indeed it just do a simple task – load the processed information from XML file and visualize based on the appropriate input sequences.

A. Tracking and Mapping components

The tracking and mapping components runs at ~5 fps on 624-352 video resolution. If a user is willing to choose a sole option either player tracking or penalty/circle center matching, its speed could be increased at ~15 fps on the same image scale.

As shown in figure 7, soccer players are tracked effectively for long-view shots, as well as the projection of five players who are marked by yellow arrows are represented on the field model. Figure 7a shows the results of tracking players where each colour of the rectangle depicts a tracked player. By solving the homogrphy matrix including the process of building field model, the trajectories of players could be displayed as in figure 7b.

The quantitative issues of our tracking and mapping components along with those obtained from previous publications are presented in Table I. From an overall perspective, the tennis tracking systems [16][27] achieves robustly the object positions mainly due to the court-view restriction where the playing time is totally captured by a single moving camera, as well as the limitation of the surrounding area. Moreover, the tennis video has a uniform court colour and

a limited player number (max = 4), this allow to detect players by searching windows and background filers.

TABLE I. COMPARISON WITH OTHERTRACKING FRAMEWORKS

Approach	Year	Domain	Tracked Players	Camera
Iwase [26]	2004	soccer	multiple	multiple
Han [16]	2006	tennis	4 (max)	single
Jiang [27]	2009	tennis	4 (max)	single
Ours		soccer	multiple	single

In contrast, it is much more difficult to detect track players with a single moving camera in soccer domain. In fact, there are many cameras used in a soccer match, but we can only watch one view at a time. Our tracking method detect and track player on long-views which is corresponding to using a single camera. Of course, we cannot estimate the player location on real-world model when players come in the regions not including penalty areas or circle center. To overcome this drawback, Iwase et al. [26] proposed an method of tracking players using multiple views, with 15 cameras the tracjectory of players is obtained by collecting the features of all cameras. However, we do not have full images from all cameras in real broadcast videos, and the camera locations are not disposed as we wish. Futhermore, the storage for 15 cameras is extremely huge.

B. OCR module

The average speed of the OCR module is 120 fps. In order to confirm our result, we focus on the information provided in the official website of 2010 FiFa World Cup South Africa [25]. Our method obtains robust results for text event detection.

TABLE II. EVENT DETECTION RESULTS

Video	Event	Occurrence	Precision	Recall
Argentina vs. Mexico	Subs.	5	100%	100%
	Cards	1	100%	100%
	Goals	4	100%	100%
Brazil vs. Chile	Subs.	6	100%	100%
	Cards	5	100%	100%
	Goals	3	100%	100%
Brazil vs. Portugal	Subs.	6	100%	100%
	Cards	7	100%	100%
	Goals	0	100%	100%
Chile vs. Spain	Subs.	5	100%	100%
	Cards	2	100%	100%
	Goals	3	100%	100%
England vs. USA	Subs.	5	100%	100%
	Cards	6	100%	100%
	Goals	2	100%	100%
Germany vs. Australia	Subs.	6	100%	100%
	Cards	6	100%	100%
	Goals	4	100%	100%
Serbia vs. Ghana	Subs.	6	100%	100%
	Cards	4	100%	100%
	Goals	1	100%	100%

Please note that: Subs. is the abbreviation Substitutes, and Cards means both yellow cards and red ones, we consider second yellow card or red card as one card.

In table 2, we gain the perfect result with 100% for seven matches in the 2011 FiFa World Cup Championship. Comparing this module to other state-of-the art methods, our method is much more simple and works very fast. In the retrieval system, OCR module could be flexibly changed by an another open source if needed. But at this time, as we assess experimentally this module, it still works well.

C. Query and Visualization

In overall, the proposed framework supports users to access the content of soccer videos efficiently via several functions:

- Shot views: users can query specific shot types that they need among long-views, short-views, close-up views, out-of-field views.
- Player locations: for selected long-view shots, there are two options: (1) if viewers only want to separate the players into two teams, players will be located by bounding retangles whose colours (red or blue) mark as their team. (2) if viewers want to know individual player positions, a player also bounded by a rectangle whose colour mark as a single one.
- Player mapping: the system provide a virtual ground image (aka field model) in which user can analyze team strategy.
- Events: with OCR module, events such as substitutes, cards, goals occur during a soccer game.

VI. CONCLUSION

We have presented a method for soccer information retrieval on the broadcast data. By providing an offline process, the system can run real-time when user query. The trajectory of players on planar field model supports for a high semantic analysis in soccer videos.

In the future, we would like to improve this method with code optimization so as to achieve a real time system. In OCR module, we also would like to build on-screen box detector in order to make the current module more completely automatic in querying text-based soccer events.

ACKNOWLEDGMENT

We would like to express our gratefulness to Thang Dinh, Bac Vo, Vu Hoang and the Computer Vision research group at the University of Science, VNU-HCMC in helping throughout the development of this project.

REFERENCES

[1] J. Zhou, and J. Hoang, "Real time robust human detection and tracking system," IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[2] C. Stauffer, W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," In: Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Ft. Collins, USA, 2246–2252 (1999).

[3] K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Background modeling and subtraction by codebook construction," In: IEEE International Conference on Image Processing (ICIP), 2004

[4] A. Elgammal, D. Harrwood, L. S. Davis, "Non-parametric model for background subtraction," In: Proc. European Conference on Computer Vision, Dublin, Ireland, 2000.

[5] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," In IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[6] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 886–893, 2005.

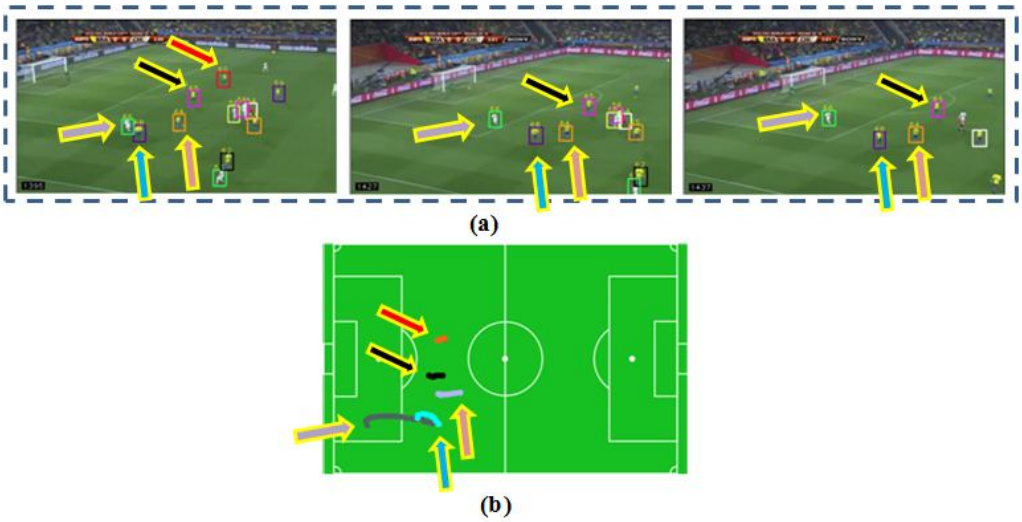

Figure 7. Trajectory of players: (a) tracking results on long-view shots, (b) player positions on field model. Each yellow arrow points out the corresponding player position on frame as well as on field model.

[7] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: people detection and articulated pose estimation," In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, 2009, 1014–1021.

[8] C. Yang, R. Duraiswami, and L. S. Davis, "Fast multiple object tracking via a hierarchical particle filter," In: International Conference on Computer Vision, 2005.

[9] Y. Li, H. Ai, T. Yamashita, S. Lao, M. Kawade, "Tracking in Low Frame Rate Video: A cascade particle filter with discriminative observers of different lifespans," In: IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[10] M. Isard and A. Blake, "Condensation -- conditional density propagation for visual tracking," In International Journal of Computer Vision. 5--28, 1998.

[11] Z. Kalal, J. Matas, K. P-N Mikolajczyk, "Learning: bootstrapping binary classifiers by structural constraints," In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[12] R.T. Collins, "Mean-Shift Blob Tracking through Scale Space," In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. vol. 2, 234-240, 2003.

[13] N. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," In: Proc. Seventh European Conf. Computer Vision, vol. 4, 373-387, 2002.

[14] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association detection responses," European Conf. Computer Vision, 2008.

[15] D. Farin and S. Krabbe and W. Effelsberg and P.H.N. de With, "Robust camera calibration for sport videos using court models," In: SPIE Storage and Retrieval Methods and Applications for Multimedia, 2004.

[16] J. Han, D. Farin and P.H.N. de With, "Multi-level analysis of sports video sequences," In: SPIE Conference on Multimedia Content Analysis, Management, and Retrieval, pp. 1--12. San Jose, USA (2006).

[17] B. Dang, A. Tran, T. Dinh, T. Dinh, "A Real Time Player Tracking System for Broadcast Tennis Video," In the 2nd Asian Conference on Intelligent Information and Database Systems, Hue, Vietnam, 2010.

[18] X. Yu, Q. Tian, and K. W. Wan, "A novel ball detection framework for real soccer video," In Proc. ICME 2003, Vol II, 265-268.

[19] X. Yu, C. Xu, Q. Tian, and H. W. Leong, "A ball tracking framework for broadcast soccer video," In Proc. ICME 2003, Vol II, 273-276.

[20] Q. Tran, A. Tran, T. Dinh, and D. Duong, "Long-view player detection framework algorithm in broadcast soccer videos," In 7th International Conference on Intelligent Computing, Zhengzhou, China, 2011.

[21] Q. Tran, B. Vo, T. Dinh and D. Duong, "Automatic Player Detection, Tracking and Mapping to Field Model for Broadcast Soccer Videos," In the 9th International Conference on Advances in Mobile Computing & Multimedia (MoMM-2011), Ho Chi Minh City, Vietnam.

[22] J. Clarke, S. Carlsson, A. Zisserman, "Detecting and tracking linear features efficiently," In: Proc 7th British Machine Vision Conference, Edinburgh, 415--424, 1996.

[23] F. Wang, L. Sun, B. Yang, and S. Yang, "Fast arc detection algorithm for play field registration in soccer video mining," In: Systems, Man, and Cybernetics Conference Proceedings. IEEE, October 2006, 4932–4936.

[24] D. Chen, J. Odobez, H. Bourlard, "Text detection and recognition in images and video frames," IDIAP, 2003

[25] The Official Website of the 2010 FIFA World Cup South Africa. www.fifa.com/worldcup/archive/southafrica2010/index.html

[26] S. Iwase, H. Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images," In: ICPR. (2004) 751-754.

[27] Y. C. Jiang, K.T. Lai, C.H. Hsieh, M.F. Lai, "Player detection and tracking in broadcast tennis video," In: Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology, pp. 759--770. Tokyo, Japan (2009).

[28] A. Ekin, A.M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Trans. on Image Processing, Vol. 12, pp.796-807, (2003).