RACKET: Real-time Autonomous Computation of Kinematic Elements in Tennis

Julien Pansiot, Ahmed Elsaify, Benny Lo and Guang-Zhong Yang
Imperial College London
SW7 2AZ London, UK
{jpansiot|aelsaify|benlo|gzy}@doc.ic.ac.uk

Abstract

This paper proposes the use of a Visual Sensor Network (VSN) for tracking the motion of tennis players on court in real-time. The proposed autonomous and wireless VSN nodes are miniaturised and powered by battery, making them ideally suited for monitoring training sessions and matches at any location. With the proposed framework, the player is tracked in the image plane using a statistical background model and efficient on-node processing. To improve the usability of the system, the normal markings on the tennis court are used as a calibration grid and the calibration algorithm is implemented with on-node processing. The node further incorporates an HTTP server to simplify transmission and interrogation of the on-node processing results by using mobile devices. The proposed system is capable of tracking a tennis player at 10 to 15 frames per seconds. Multiple nodes are deployed simultaneously either to track several players or to enhance the tracking accuracy of a single player. Features related to motion and game tactics are used to guide the training sessions and refine player tactics.

1. Introduction

For elite tennis players, subtle physical adjustments or mental attitude changes can significantly affect performance outcomes of major tournaments [5, 11]. For this reason, detailed analysis of biomotion and cross-court movement has become an important training tool.

In order to continuously refine training strategies, players and coaches rely on key information recorded over the training sessions and matches. Logbooks are commonly used to monitor progressive changes in training [11], which provide general high-level information about the nature and outcome of the training sessions. Video recordings during training and matches are becoming increasingly popular recently [15]. Current advances in information technology allow large amounts of video data to be collected and stored easily, but retrieving useful information from these datasets through post-processing remains a time consuming task.

In order to provide the players and coaches with ready-to-use biomechanical parameters, marker-based motion tracking has often been used. However, most commercial systems involve a large set of wearable markers or sensors [16, 18], and thus are obtrusive. They are therefore typically used to record a snapshot of the player's technique but not used on a regular basis.

Video-based monitoring is a tool of choice for detailed sport motion analysis as it does not affect the players, thus providing more detailed measurements for both training and competition. However, plain video recordings do not provide direct metrics to the players or coaches. Further image processing is required to derive information such as location, velocity or detailed posture parameters. The purpose of this paper is to present a VSN system that would combine the ease of video recording with real-time extraction of biomotion features comparable to marker-based systems. By using multiple VSN nodes, the system is able to provide extended coverage with improved accuracy. Wireless communication simplifies the installation process, particularly during tournaments, while on-board processing reduces communication requirements.

The rest of this paper is organised as follows. Section 2 provides an overview of the related work. Section 3 introduces the VSN platform used in this work. Section 4 details the configuration and calibration of the system. Section 5 presents the player tracking algorithm and Section 6 provides practical results before concluding in Section 7.

2. Related work

In this paper, three levels of information that can be extracted from a video sequence are distinguished: the player's position and velocity on the court, the player's technique (stroke and short term tactics) and game strategy. The purpose of this paper is to focus mainly on positional tracking. Early work on tennis player tracking was performed by Sudhir *et al*. [13] and Pingali *et al*. [12]. The system proposed by Sudhir *et al*. encompassed court line detection for camera calibration, player tracking, as well as high-level

feature detections. Court detection and camera calibration are fundamental features for tennis tracking systems. It allows the derivation of the position of the player on the court from apparent two-dimensional (2D) image features. Pingali *et al*. used four cameras to track the player as well as the ball, and introduced the concept of the occupancy map.

Bloom and Bradley [3] later carried out similar work and extended it with stroke recognition. The sound track was also used to detect the exact time of ball contact and derived the player's apparent skeleton to infer the stroke being played. Wang and Parameswaran [17] proposed to classify the 58 winning patterns recommended by the US Tennis Association (USTA) [14].

Thus far, most of the tennis-related computer-vision studies have focused on tennis ball trajectory tracking. Commercial systems such as Hawkeye [7] are already mature enough to be used routinely in major tournaments.

3. VSN node design

The mobile VSN node developed at our laboratory is a self-contained module composed of three stackable boards and a battery. The main board embeds a 500MHz Analog Devices Blackfin BF537 Processor [1], 256Mb SDRAM, 32Mb SPI Flash. The camera board is based on an Omnivision OV9655 1.3 megapixel camera [10] with interchangeable lenses including a range of focal lengths. Wireless communication is provided by a Lantronix Matchport [8] WLAN 802.11g/b Wi-Fi board.

The physical dimension of the module is $25\times45\times45$mm, with a total weight of 80.5g. Its total power consumption is 220mA in fully active mode (*i.e*., during concurrent frames acquisition, compression and wireless communication). A detailed breakdown of the power budget based on using a 2.8Ah battery includes: main processor board: 147mA, wireless communication board: 55mA, camera board: 21mA. In sleep mode, the total power consumption drops to 10mA. The module components are illustrated in Figure 1.

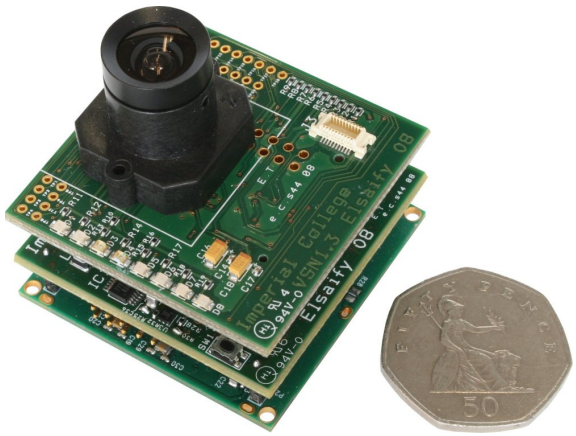

Figure 1. The proposed self-contained VSN module.

4. Node calibration

In order to derive relatively accurate three-dimensional (3D) position of the player on the court from 2D image sequences, it is necessary to calibrate the node *in situ* to determine extrinsic camera parameters. To avoid the use of an extra calibration grid or fiducials, the markings on the tennis court are used. This operation is conducted in two steps involving corner detection and actual camera calibration.

4.0.1 Tennis court corner detection

The operator is asked to point the VSN node near the centre of the *middle T* of a tennis court. After acquiring a still image from the VSN node, the relevant tennis court lines intersections are detected by using the following method.

In order to determine the local tennis court marking topology at a given point, a polar intensity histogram is built around this point. Thus the histogram represents the average image intensity for a given angular sector in the vicinity of the considered point. The intensity histogram can be robustly and quickly segmented, with local maxima denoting the tennis court line (as it is brighter than the court surface). The method has proved to be robust enough to deal with thin lines of sub-pixel width. Depending on the number of maxima, further topological assumptions are made, such as line junction detection.

A fully automated system has been designed that can detect the relevant tennis court corners with on-node processing. Starting with the *middle T*, the algorithm follows the service line in both directions until the intersections with the *side lines* are found. The *side lines* are then followed to find their intersections with the *baseline* at the back of the court, as illustrated in Figure 2. This process is relatively immune to noise and can be executed in less than a millisecond on the VSN node. The detected corners are then used to calibrate the camera.

4.0.2 Extrinsic camera calibration

The camera calibration aims at determining the camera position (x, y, z) and its orientation defined by the Euler angles (α, β, γ) with respect to the tennis court. A gradient descent algorithm optimises the camera position and orientation by reprojecting the tennis court corners and matching them with the actual court geometry. The camera position and orientation are then retrieved. Random seeds can be used to prevent the gradient descent algorithm from being locked in local minima. Finally, the back projection matrix of the camera is computed and the node is eventually

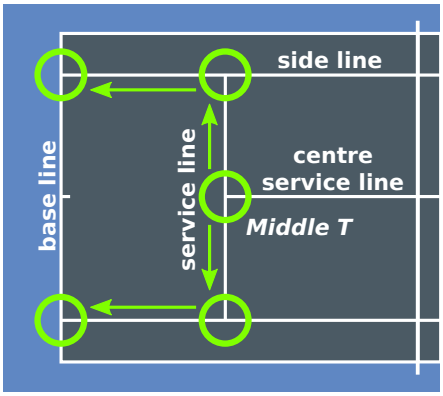

Figure 2. Tennis court corners automatically detected on-node for camera calibration.

ready for tracking with aligned image and world coordinate systems.

An optimised version of this algorithm has been ported onto the VSN node and the calibration process takes about 10 to 30 seconds. With this approach, it is possible to rely entirely on the embedded software to detect the corners and calibrate the VSN camera. Therefore, the VSN node can be used in a completely autonomous fashion.

4.0.3 Camera distortion correction

Radial distortion is a significant problem when using small lenses with relatively short focal lengths. An optional distortion correction step is therefore provided in the proposed framework. In this paper, a polynomial radial camera distortion model [4] was used:

$$\mathbf{d_N} = \mathbf{u}\left(1+\sum_{i=1}^{N} k_{2i+1}\|\mathbf{u}\|^{2i}\right) \quad (1)$$

$$\mathbf{d_1} = \mathbf{u}\left(1+k_3\|\mathbf{u}\|^2\right) \quad (2)$$

where $\mathbf{d}$ is the distorted point in the normalised coordinates system and $\mathbf{u}$ in the undistorted, pinhole projection point. The reverse equation can be approximated at the second order:

$$\mathbf{u} \approx \frac{\mathbf{d}}{1+k_3\left(\frac{\|\mathbf{d}\|}{1+k_3\|\mathbf{d}\|^2}\right)^2} \quad (3)$$

In the experiments conducted in this paper, the average error of this step was found to be 0.75 pixels, with maximum error being 3 pixels at far-end corners. A more accurate solution can be achieved by recursive application of the above equation.

For VSN programming, floating point operation is computationally expensive. For this reason, a look-up table providing pixel displacements is first created and used subsequently. This solution proves to be efficient but has two main weaknesses. The overall processing time is increased by about 20% and the re-sampling artefact is significant.

To circumvent this problem, a novel hybrid approach is proposed. The look-up table is first used to correct the distortion on the whole image during the calibration phase. At this stage, the computation speed is not an issue and re-sampling artefact only has a limited impact. During the actual tracking phase, image processing is performed on the distorted image, at no extra computational cost and without artefacts. Only player's feet position in the image space is undistorted using a floating point method before final mapping back onto the tennis court coordinate system, which has virtually no effect on the overall performance.

5. Tennis player tracking

5.1. On-node background segmentation

In order to separate the player from the rest of the scene, background subtraction is performed. Pixel-based segmentation methods using statistical distribution of the background colour are employed in this study. The colour model used in this paper is based on Gaussian Mixture Models (GMM) [9].

Since the processing power of the VSN node is limited whereas GMM requires the use of extensive floating point operation (not supported in hardware by the Blackfin processor), it is only possible to use a single Gaussian for the background colour model for real-time operation, which is tantamount to a mean and variance model. It has been shown that the node processing power is sufficient to enable on-node, real-time background segmentation at a resolution of 320×256 pixels (quarter SXGA).

Alternative methods such as histogram-based methods [6] or multi-modal mean (MM) [2] could potentially have been used for this purpose in order to model a multi-modal background, but they are more demanding. A very simplified fixed-point implementation of MM would require roughly 8 times more multiplications (4 cells × 2 multiplications for the fixed-point). A histogram-based method would require a considerably larger amount of memory to store the model (320 × 256 pixels × 10 bins × 3 colours = 2,400KB).

Further optimisation is necessary to achieve truly real-time performance. First, the background model is only updated every 20 frames. Secondly, a low-resolution segmentation is performed beforehand to determine the Region Of Interest (ROI). Full resolution segmentation and morphological filtering (erosion-dilation) are then performed only within the ROI. This hierarchical method provides a substantial speed gain, with the segmentation step completed in 22ms.

After segmentation, several low-level features can be computed from the binary blob image including centre, the

Axis-Aligned Bounding Box (AABB), and the eigenvectors of the blob determining its global orientation.

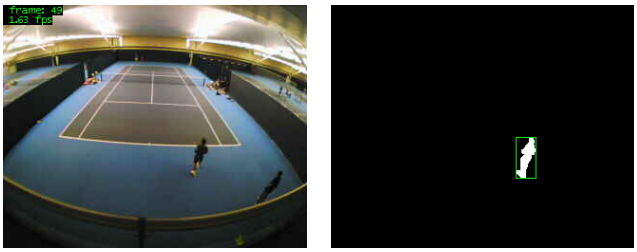

Figure 3. Left: original image as captured by the VSN node camera using a wide angle lens. Right: on-node binary blob segmentation and AABB computation.

5.2. Monocular player tracking

In order to track the player movement, background segmentation is performed and the bounding box around the player is computed. The position of the feet is then extracted from the bounding box in the 2D image space and back-projected into the real-world 3D coordinates by using the calibration matrix. This under-constrained problem is solved by assuming that the players feet are touching the ground (ground plane constraint).

Due to the highly non-linear and hardly predictable player's motion, temporal filtering of the trajectory using a prediction-correction tracking algorithm such as Kalman filter improves only marginally the system accuracy and was therefore discarded.

5.3. Multiview player tracking

Whilst the ground plane constraint holds most of the time, it can fail and lead to inaccurate results. This happens typically in two cases, *i.e.*, if the player jumps up or if for some reason the legs are incorrectly segmented. In either case, the lowest visible part of the player is not in direct contact with the ground and the constraint no longer holds.

One of the simplest solutions for overcoming this problem is to add a second VSN node. By calculating the intersection of the line of sights of the VSN node on the ground plane, a more robust position can be derived. This solution relaxes the aforementioned ground plane constraint and is particularly resilient to inaccuracies in the vertical axis of the image space.

Given the relatively low cost of the proposed VSN nodes, it would therefore be possible to consider setting up a larger number of them in order to increase the tracking coverage and accuracy. However promising, this approach has not been studied into more details, as the main drive behind this particular application is in its compactness and ease of use.

5.4. Computational load analysis

In order to demonstrate the value of on-node processing, Table 1 compares the bandwidth requirements for a centralised processing scheme requiring video or feature stream communication.

Image encoding (320×256 pixels)	Size (bytes)	Frame rate (frames/s)
Raw RGB	245,760	0.4
Raw YUV 422	163,840	0.6
JPEG (high quality)	24,650	2.6
JPEG (medium quality)	5,960	6.4
JPEG (low quality)	3,820	8.7
Binary blob (raw)	10,240	5.6
Binary blob (run-length)	847	11.3
Features only	40	14.7

Table 1. VSN node image output size and frame rate comparison. It can be observed that on the proposed platform, on-node processing allows higher frame-rate by reducing dramatically the time spent on communication.

Computational loading analysis was performed during the development and optimisation phases of the system. Figure 4 summarises the distribution of the image processing tasks previously described on the processor. It is evident that a combined use of low-resolution pre-segmentation and AABB reduces the overall computation time dramatically.

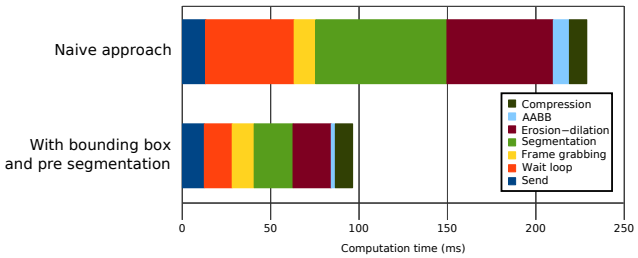

Figure 4. Software optimisation for on-node processing showing the computational loading of the main operations involved in the player tracking.

6. Practical applications and results

6.1. Experimental setup

An upper concourse at the back of the tennis courts is used for assessing the practical deployment of the proposed system. The VSN nodes are placed at a typical height of 6 meters, 20 meters away from the centre of the court. Ideally, the sensors would be placed at either end of the court, but because only one concourse was accessible, an asymmetric configuration with two different focal lengths was used.

Figure 5 illustrates the difference of perspective projection between the near- and far-side tracking. The vertical resolution on the far side is bound to be lower, even when using a longer focal length sensor, as the angle between the optical axis of the camera and the court plane is significantly lower.

Figure 5. Left: near side of the court as seen by the VSN node with a 3.6mm focal length optic. Right: far side as seen with a 6.5mm focal length.

Furthermore, it should be noted that the proposed system was mostly tested indoor, where the lighting is relatively consistent over time.

6.2. User queries

In order to provide efficient access to the VSN node, a software environment has been developed for real-time on-node data interrogation. To facilitate its use by non-technical users, a micro web server has also been implemented on the VSN node. This allows any computer or hand held devices (such as an Apple iPhone, as illustrated in Figure 6) to retrieve live data from the VSN node. In order to reduce the bandwidth usage, a lightweight AJAX framework is used.

Figure 6. Tennis player tracking interface running on an Apple iPod Touch.

6.3. Strategy metrics

The on-node real-time tracking was trialled during a Junior's Masters Tournament, as well as two of the matches of GB Davis Cup Team play-offs. The positions of the players were also archived in real-time on a PC for further analysis. During some of the matches, the game was also manually annotated to provide ground-truth data for comparison.

6.3.1 Direct feature representations

The fundamental kinetic parameters are provided in real-time or during playbacks. These include the total distance covered and the current speed and acceleration. Several representations of the trajectory were made available to the coaches. The actual trajectory on the court over an arbitrary length of time can be readily played back during the game. Some trajectory patterns can be easily observed. This is particularly useful during training "drills", when the player is asked to repeat the same movement a number of times.

After data collection, it is possible to derive a court occupancy map as illustrated in Figure 7. This map represents the proportion of time spend by the player at different locations on the court, which can be used to infer general tactics during the game.

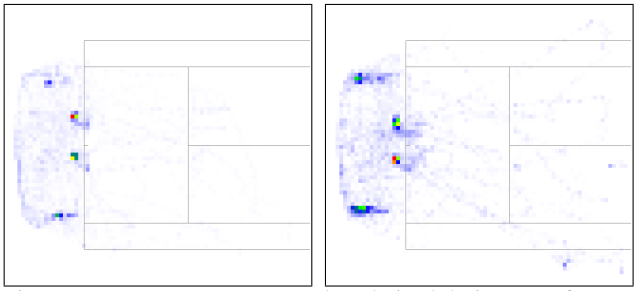

Figure 7. Two court occupancy plots derived during a set for two competitors. Serve and return locations are clearly visible. The right graph shows more mobility and more variability on the serve return placement.

6.3.2 Player tactics-related metrics

In general, the tracked features described earlier can indicate either instantaneous or global motion characterisation of a player. In either case, the outcome remains a direct representation of the recorded values. Higher-level metrics can be constructed by aggregation and analysis of patterns in the original feature set over time.

Assuming that both players have been tracked and their shots have been manually annotated, it is possible to segment the games accordingly. Indeed, the serves can be identified automatically using low-level cues such as player position and inter-shot times. For each game, basic statistics can be derived, such as the number of shots, game duration, and distance covered by each player.

An attempt to recognise some of the winning patterns described in [14] has been carried out. Whilst this has been performed initially by Wang and Parameswaran [17], they

relied on ball tracking, whereas the proposed system only tracks the player. This raises certain difficulties as the aforementioned winning patterns are defined in terms of ball ground contact.

For pattern recognition purposes, the tennis court has been divided into nine zones. Due to the limited annotated data available, a zone-based segmentation was employed. The focus was set on the serve and fifteen winning patterns and two variations related to serve and return were considered. However, due to game style adopted by the players, only four patterns have been frequently recognised: *return deep cross-court* (7), *return deep down the middle* (10, 11) and *return low at the server's feet* (17). An example of such pattern is shown in Figure 8.

6.4. Tracking accuracy validation

In order to assess the overall accuracy of the proposed system, its spatial accuracy on the nearest side of the court was evaluated by using a ground-truth grid. A metric grid was carefully marked on the ground with a masking tape. Two players were then asked to move from corner to corner on the grid whilst a VSN node was tracking and transmitting their positions to a computer for recording. This experiment was performed twice to ensure a complete coverage of the court, and the results are shown in Figure 9.

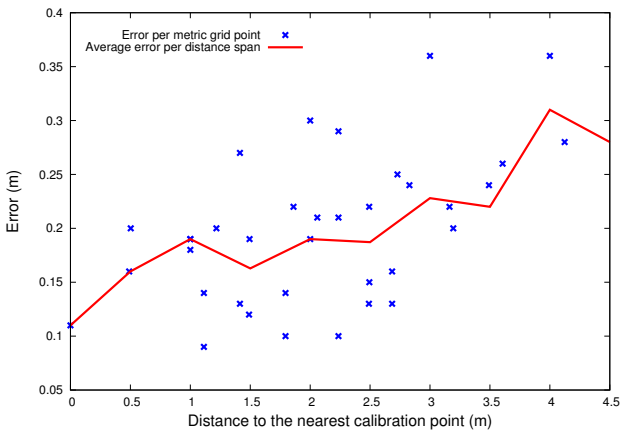

Figure 9. Error due to distortion: when the player stands far away from all calibration points, the camera distortion leads to an erroneous position estimation.

Results have shown that the overall average accuracy of the system is 20cm, with the minimum and the maximum at 9cm and 36cm, respectively. This error level appears suitable for pattern of play recognition. Three main sources of inaccuracy are observed during this experiment.

Distortion The experiment was performed without distortion correction, which affected the accuracy near the serve point, as it is far away from the calibration points.

Incorrect binary segmentation Typically, if the player wears a colour clothing similar to the court, the tracker will be less sensitive to motion.

Low spatial resolution During this experiment, the horizontal resolution near the service line was 5cm per pixel. This means that the legs of a player will appear no thicker than two pixels.

7. Conclusion and future work

In this paper, it is has been shown that the motion of the players can be monitored during competitive events without interfering with the game. It has also been demonstrated the value of the VSN nodes in terms of fast deployment for practical applications. The miniaturised size of the VSN nodes combined with their low-power consumption and wireless communication, makes them particularly attractive for field-sports. As the majority of the training sessions for top players are not fixed at a single location, the proposed system will allow for coaches and players to carry out detailed analysis *in situ* and when it matters.

The full implementation of the VSN node calibration and player tracking on the node means that no specific software is required on the user side, and it is therefore possible to visualise the motion patterns on hand-held devices. The archived data also permits detailed off-line analysis. The match or training session can be replayed, and strategic patterns can be automatically identified.

Further optimisation might be necessary in order to perform player tracking with higher spatial and/or temporal resolution. Automated annotation of the shots could be performed based either on ball tracking or by incorporation of other sensing modalities, such as sound. It would also be useful to extend the current system to include explicit 3D reconstruction to further enhance the overall accuracy of the system.

Acknowledgements

The authors would like to thank Karl Cooke from the Lawn Tennis Association (LTA), for facilitating numerous experiments on high-profile tennis matches and his insights on the coaches and players needs.

References

[1] Analog Devices. Blackfin BF537. http://www.analog.com.

[2] S. Apewokin, B. Valentine, D. Forsthoefel, L. Wills, S. Wills, and A. Gentile. Embedded real-time surveillance using multimodal mean background modeling. *Embedded Computer Vision*, pages 163–175, 2009.

[3] T. Bloom. Player tracking and stroke recognition in tennis video. In *Proceddings of WDIC*, pages 93–97, 2003.

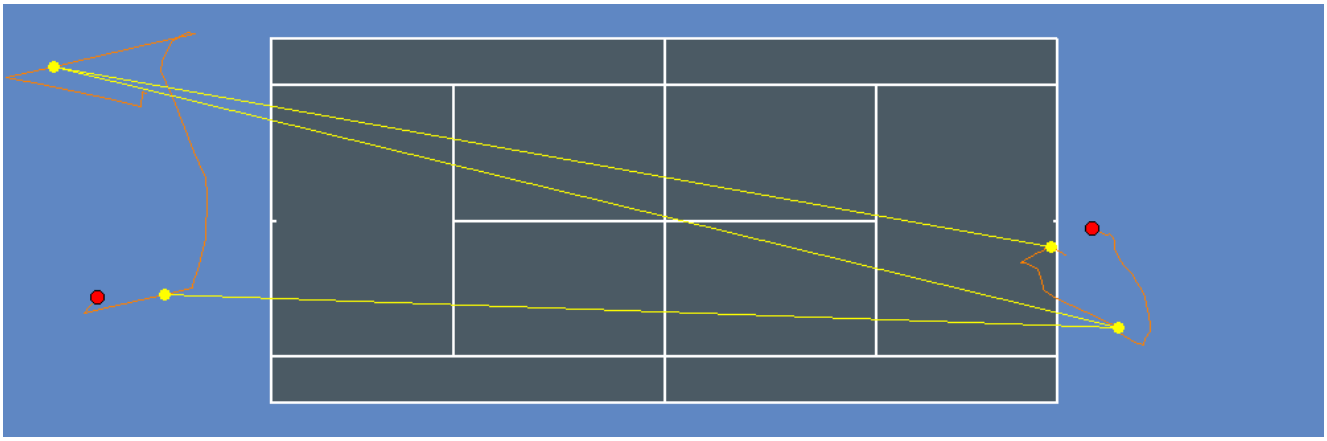

Figure 8. Tennis winning pattern recognition: “Against a wide serve - return deep cross-court” [14]. Note that the lines only represent the ball exchanges, not its actual trajectory.

[4] D. Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37(8):855–866, 1971.

[5] D. Gould, R. Medbery, N. Damarjian, and L. Lauer. A survey of mental skills training knowledge, opinions, and practices of junior tennis coaches. *Journal of Applied Sport Psychology*, 11:28–50, 1999.

[6] D. Gutchess, M. T. C, E. Cohen-solal, D. Lyons, and A. K. Jain. A background model initialization algorithm for video surveillance. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 733–740, 2001.

[7] Hawkeye. http://www.hawkeyeinnovations.co.uk.

[8] Lantronix. Matchport. http://www.lantronix.com.

[9] D.-S. Lee. Effective gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(5):827–832, 2005.

[10] Omnivision. OV9655. http://www.ovt.com.

[11] C. Petersen and N. Nittinger. Fit to play: making better players, on & off court. *Medicine and Science in Tennis*, 9(1):20–21, 2004.

[12] G. Pingali, Y. Jean, and I. Carlbom. Real time tracking for enhanced tennis broadcasts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 260–265, Jun 1998.

[13] G. Sudhir, J. Lee, and A. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proceedings of the IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 81–90, Jan 1998.

[14] United States Tennis Association (USTA). *Tennis Tactics: Winning Patterns of Play*. Human Kinetics, 1996.

[15] P. C. W. Van Wieringen, H. H. Emmen, R. J. Bootsma, M. Hoogesteger, and H. T. A. Whiting. The effect of video-feedback on the learning of the tennis service by intermediate players. *Journal of Sports Sciences*, 7:153–162, 1989.

[16] VICON. Motion capture systems. http://www.vicon.com.

[17] J. Wang and N. Parameswaran. Analyzing tennis tactics from broadcasting tennis video clips. In *Proceedings of the 11th International Multimedia Modelling Conference*, pages 102–106, Jan. 2005.

[18] Xsens. Moven. http://www.moven.com.