Recognition of Badminton Strokes Using Dense Trajectories

Sameera Ramasinghe, K. G. Manosha Chathuramali, Ranga Rodrigo
Department of Electronic and Telecommunication Engineering
University of Moratuwa
Sri Lanka
Email: samramasinghe@gmail.com, manosha@ent.mrt.ac.lk, ranga@uom.lk

***Abstract*—Automatic stroke recognition of badminton video footages plays an important role in the process of analyzing players and building up statistics. Yet recognizing activities from broadcast videos is a challenging task due to person dependant body postures and blurring of the fast moving body parts. We propose a robust and an accurate approach for badminton stroke recognition using dense trajectories and trajectory aligned HOG features which are calculated inside local bounding boxes around players. A four-class SVM classifier is then used to classify badminton strokes to be either smash, forehand, backhand or other. This approach is robust to noisy backgrounds and provides accurate results for low resolution broadcast videos. Our experiments also reveal that this approach needs relatively fewer training samples for accurate recognition of strokes compared to existing approaches.**

***Keywords*—Badminton stroke recognition, action recognition, dense trajectories, HOG, SVM**

I. Introduction

Automatic action classification of low resolution videos is a challenging task. In the domain of automatic action classification of sports videos, this task becomes comparatively simple due to the predictability of action categories that could occur in a sport. But it may entail unique challenges of its own due to several reasons. For example, body postures of the players while the same stroke is being played might be different from player to player, and, even for a single player, it might be different for two occasions. In general, players move rapidly throughout the video. Therefore limbs and sport equipments become blurred most of the time. Inability to extract useful features accurately out of blurry foreground objects makes it a cumbersome task to identify the strokes and events. A broadcast video is a post-edited video distributed through multimedia channels such as television, Internet, and is generally of poor quality. Therefore, in broadcast videos, the process becomes even more difficult.

Overall, the level of action classification can be mainly divided in to two categories, higher level and lower level. In higher level action classification, the focus is more on obtaining an output which is more semantically meaningful to the viewer than in lower level action classification. An example for a higher level output could be "the game had a rally for 4 minutes and then player 1 scored while player 2 was near the net". Although the higher level action classification is more semantically meaningful to a general viewer, it does little work on empowering the process of building up statistics and player analysis. For an example, in building up statistics, it is more useful to have the exact number of strokes played by each player with a timeline. Lower level action classification, on the other hand, if it is able to identify low level actions accurately such as individual strokes, is much more useful in that context. Therefore, our focus in this paper is on building up an accurate low level action classification system for identifying badminton strokes from low resolution broadcast videos, which would provide a platform for automatic analysis of players and generating statistics.

Similar to the level of action classification, existing work on automatic sports video annotation can be broadly categorized in to 3 categories based on the algorithmic approach: 1) appearance based methods (Connaghan *et al*. [1], Kijak *et al*. [2], Conaire *et al*. [3], Bloom *et al*. [4]), 2) combining player, ball locations and domain knowledge for reasoning (Sudhir *et al*. [5], Miyamori *et al*. [6], Gong *et al*. [7], Pingali *et al*. [8]), and 3) motion based feature methods (Zhu *et al*. [9]). Since our focus on this paper is automatic recognition of badminton strokes, it is useful to have a brief insight in to some of the work done on automatic tennis stroke recognition in detail, which is a similar application.

First category, appearance based methods, is the most popular and traditional approach. Connaghan *et al*. [1] describe a method for action classification in tennis videos, where the players silhouettes are extracted via background subtraction. These silhouettes are then further analyzed and hu moments [10] of binarized player blobs and mahalanobis distance [11] between them are used for classification of strokes in to three classes: serve, backhand, and forehand. But they use recorded sessions in controlled environments in their work where background clutter and crowd movements are not present. This makes it hardly applicable to noisy broadcast videos. Kijak *et al*. [2] merge audio and video information to classify actions in to four classes: serve, rally, replay, and break. In the process of extracting visual information, they use an appearance based method, where they measure the visual similarity between frames using a defined weighted function of the spatial coherency, distance between dominant color vectors, and activities. Conaire *et al*. [3] propose fusion of both

978-1-4799-4598-6/14/$31.00

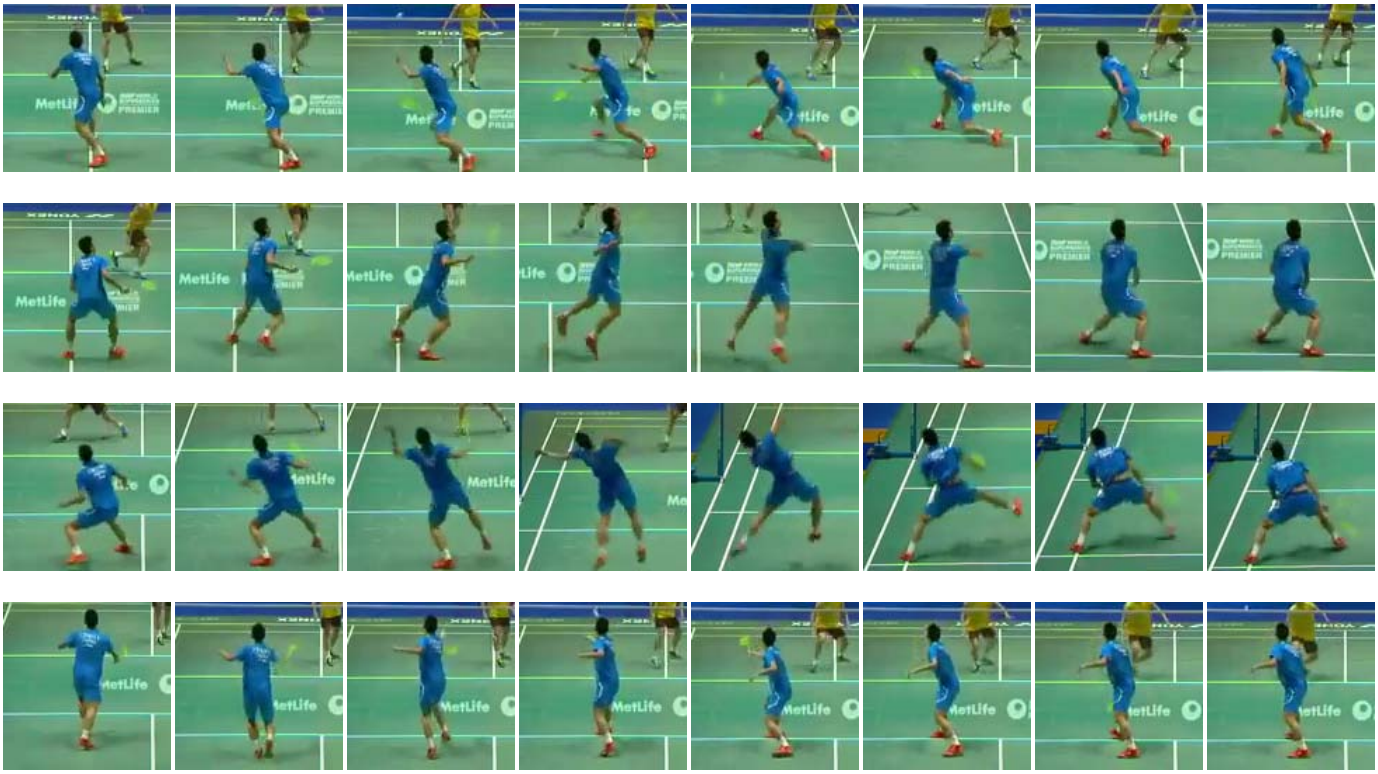

Fig. 1: Example image sequences for each class. Top row: backhand, second row: forehand, third row: smash, last row: other

sensor information and visual information in order to classify tennis strokes to be either forehand, backhand, or serve. For extracting visual features, player silhouettes are obtained via background subtraction. Moreover, these silhouettes are then divided in to 16-pie radial segments centered on player centroids, and a feature vector is created storing the largest distance from the player centroid to the silhouette boundary in each segment. This feature vector is then used to classify strokes. Bloom *et al*. [4] also present an appearance based approach which is similar to [1], but slightly different: They classify tennis strokes into four main categories: forehand ground stroke, backhand ground stroke, volley, smash, and serve. The strokes are identified by analyzing the relative position of the racket compared to the player. For extracting the racket position and the player position, they use background subtraction. In general, appearance based methods rely on extraction of silhouettes in event analysis.

Most of the work falling in to the second category use prior knowledge of the particular sport domain for reasoning, and tracking of objects as a low level information collector. Sudhir *et al*. [5] follow the process of first building up a model of the court by extracting court lines in the video. Then the position of the player with respect to the court is identified and this position information is then fed to a higher level reasoning module. Based only on these position information of the player, the reasoning module is able to recognize activities such as baseline rallies, passing shots, net-games, serve, and volley games. The work of Miyamori *et al*. [6] could be partly classified in to both first and second categories. They extract court lines to identify the player location and use player silhouettes for behavior identification. Then all the information is fed to a reasoning module to classify the action into classes: forehand stroke, backhand, volley and smashing. However an accurate stroke classification, based only on player position, without tracking the ball or shuttlecock, is an impractical task. Tracking the ball or shuttlecock is also hardly practical in broadcast videos. For low level action classification, therefore, the player body postures has to be re-analyzed to come to an accurate conclusion even for approaches of this category, as in Miyamori *et al*. [6]. This method, however, is appropriate for high-level action classification compared to other methods, as it provides position information of the player and the ball or shuttlecock, which are useful in high-level reasoning.

Using motion based features for action classification is the third category. Zhu *et al*. [9] use the information obtained from optic-flows for tennis stroke recognition. They derive a feature called *slice-based-optic flow-histograms* based on optic-flow clutter for this purpose. In their work, they classify the strokes into two categories: left-swing and right-swing. This approach gives more emphasis to motion information and motion pattern information for analyzing and recognizing actions.

In all three categories, the main drawback of applying appearance based models for broadcast badminton videos is that they mostly rely on accurate extraction of player silhouettes for action recognition. But this may be impractical in low resolution broadcast videos due to blurring and noise

Fig. 2: Upper diagonal camera angle

TABLE I: Sample set of data written to the database while manually annotating the temporal and special locations of strokes

Frame	Shot Number	CentroidX	CenroidY
32267	1	300	257
32268	1	302	259
32269	1	305	261
32270	1	299	259
...	...	...	...
32310	2	405	206
32311	2	270	190
32312	2	280	210

in the background. Also the diversity of the silhouettes due to variations of the human postures while playing the same stroke also makes it a complex task to calculate reliable features. Similarly, as a stroke consists of an array of images rather than a single image, extracting features out of the optimum image or combining the information extracted from an array of silhouettes is a challenging task based solely on blob or silhouette information. Furthermore, extracting the position of the racket is also impractical due to blurring and occlusion in real game scenarios in racket games. Fig. 1 shows several classes of actions. Due to motion blur, it is difficult to track the racket in these videos. The second method, using the prior knowledge of the playing-domain and tracking the ball or player, provides a reliable platform for higher level action classification rather than for accurate stroke recognition. Also, the methods which rely on tracking the ball is of less use in this problem domain as the ball or the shuttlecock is not visible most of the time in broadcast videos. So in this context, the best approach for automatic identification of badminton strokes is the third method, which captures the general motion pattern of the player's body in the feature extraction layer.

In this paper, we propose a robust and efficient method for automatic stroke classification of low resolution broadcast badminton videos using dense trajectories [12] based trajectory-aligned HOG features [13]. Wang *et al.* [12] apply dense trajectories for the whole video sequence and to the entire frame to gather higher level information about the video. Five trajectory aligned features, namely, trajectory descriptor, HOG, HOF, mbX, and mbY are separately used to create bag of visual words (BoVW) models and fed to an SVM classifier. SVM results are then combined to achieve the final result. In our work, however, we used dense trajectories with few alternations as applicable to our problem: We treated dense trajectories as a lower level feature rather than a higher level feature. We created dense trajectories only inside bounding boxes around players. Dense trajectories were created iteratively after every stroke inside the same video. The length of the trajectories were reduced as a single badminton stroke only extends up to maximum of around 30 frames. We only used trajectory aligned HOG features for creating a BoVW model which was effective and faster. We strictly focused on the specific problem of stroke identification and classification, and marking of the bounding boxes and the duration of each stroke was done manually in the training phase. We classify strokes in to four basic categories: forehand, backhand, smash and other. By applying our method to noisy YouTube videos, we show that our results are on par with or outperform the state-of-the-art methods.

II. Methodology

The proposed automatic badminton stroke recognition system consists of 4 basic steps: manually annotating the temporal and special locations of strokes, creating dense trajectories, creating a cluster centers for the bag-of-visual-words model, and training.

1) *Manually annotating the temporal and special locations of strokes*

Usually a broadcast video of a badminton game consists of multiple camera angles and slow motion replays. Recognition of strokes in those situations is possible using the same method, but for simplicity, in our experiment, we limited the testing only to the upper diagonal camera angle, which is the most common camera angle during a match. Upper diagonal camera angle is illustrated in Fig. 2. Since we were only interested in developing an accurate algorithm for stroke recognition, we avoided the problem of automatic human detection and automatic recognition of starting frame and ending frame of a stroke by manually marking it. We gave each stroke a unique number and stored frames corresponding to each stroke number in a database. Also, in order to mark the spatial location of the player during the shot, the coordinates of the centroid of the bounding box surrounding the player was written to the same database. The bounding boxes were of fixed size $120pixels \times 120pixels$. The size of the bounding boxes were choosen experimentally. An excerpt of data written to the database is shown in Table I.

2) *Creating dense trajectories*

After the manual annotation, we cut the bounding boxes of each frame of each stroke and fed it to the dense trajectory generating module. For the generation of dense trajectories, we used the same method as Wang *et al.* [12] with a few alternations. First a dense optical flow field was computed using the algorithm by Farnebäck [14], and points were tracked densely for 8 multiple spatial scales. To form a

trajectory, these points were connected across each frame. Trajectories tend to drift from their original locations after few frames. Therefore, We reduced the length of a trajectory to 6 frames. After 6 frames, the trajectories were destroyed and new trajectories were generated at each location. The foremost reason behind choosing the number 6 is, the need to create around five trajectories for a continuously moving point to avoid drifting, as a badminton stroke typically extends up to 30 frames in average. The descriptors were calculated around a space-time volume around the trajectory. The volume was of the size 32×32 pixels and 6 frames. This volume was then subdivided in to a $2 \times 2 \times 3$ equal sized grid where dimensions are in x-direction, y-direction, and time axis, respectively. HOG descriptors with 8 bins were then calculated inside every spatio-temporal sub volume inside the grid. A $8 \times 2 \times 2 \times 3 = 96$ dimensional descriptor vector was calculated for every space-time volume around a trajectory. We used HOG features as the trajectory aligned feature descriptor as we wanted to capture static information around the player in order to extract the information about the body shape of the player as well as the motion. The motion pattern information was also captured simultaneously as HOG features were calculated along the trajectories. A visualization of dense trajectories is shown in Fig. 3.

3) *Creating a cluster centers for the bag-of-visual-words model.*

As explained above, for every trajectory, a 96-dimensional vector was created. Out of each stroke, a bag of 96-dimensional vectors was generated where the number of vectors in the bag depends on the number of trajectories generated by the stroke. For creating the cluster heads for the BoVW model, we chose 15 strokes from each of the four categories (smash, backhand, forehand, other) from a YouTube video of the resolution 760×360 and created descriptor vectors for each of the strokes. The total number of vectors was 101,750. All the vectors were then mixed and k-means clustering was then used to identify 4000 cluster heads.

4) *Training*

Each of the bag-of-vectors of each individual stroke mentioned above, was used to create a histogram for each stroke where we used Euclidian distance between the vector and the cluster centers to assign it to a cluster center. By following this procedure for each individual stroke, a 4000-dimensional descriptor vector (histogram) for each stroke was created. We used these vectors (15 from each stroke class) for our training procedure. As training procedure, we used a four-class SVM classifier. The OpenCV implementation of the SVM classifier was used for this procedure with a linear kernel and maximum number of iterations 100. After few experiments, a termination criteria of the iterative SVM training was chosen as 1×10^{-6}. For training, 60 histograms (15 from each class) were created as described in the earlier section and fed to the SVM classifier.

III. Results and evaluation

We used two main approaches for testing our work. The strokes used for training of our system were extracted from the first half of the badminton match. First, we tested our method on the strokes extracted from the second half of the same match. Then we tested our work on strokes extracted from a completely different low resolution YouTube video, played by a different player with a different stroke style. A total of 120 strokes, 60 for each method, was tested where we chose the strokes in the order they were played in the video. We made no specific attempt to choose specific strokes.

1) *Evaluation approach 1*

In this approach we tested our system on the strokes extracted from the second half of the same match used for training. Strokes are played by the same player. After testing 60 strokes, 59 were recognized correctly while one was misidentified which resulted in 98.34% overall accuracy. The detailed description of the correctly and incorrectly recognized number of actions reported over each stroke class is shown in Table II.

TABLE II: Number of tested strokes of each class in evaluation method 1

Stroke class	Number of tested shots	errors	accuracy
Smash	10	0	100%
Forehand	21	1	95.23%
Backhand	18	0	100%
Other	11	0	100%
Total	60	1	98.34%

2) *Evaluation approach 2*

Next we tested our work on strokes extracted from a different video, played by a different player with a different stroke style. After testing 60 strokes, 56 were recognized while 4 were misidentified which resulted in 93.34% overall accuracy. The results in detail are shown in table III.

TABLE III: Number of tested strokes of each class in evaluation method 2

Stroke class	Number of tested shots	errors	accuracy
Smash	11	0	100%
Forehand	24	0	100%
Backhand	12	2	83.34%
Other	11	2	81.81%
Total	60	4	93.34%

A. Interpretation of results

Although our system was able to keep the recognition rate over ninety, it is beneficial to analyze the results in detail to recognize the reasons for misrecognized strokes. In evaluation approach 1, the misidentified shot was a forehand stroke where the body posture was very similar to a backhand shot as shown in Fig. 4a. HOG features capture information about the general shape of the body, which is close to a backhand stroke. Since

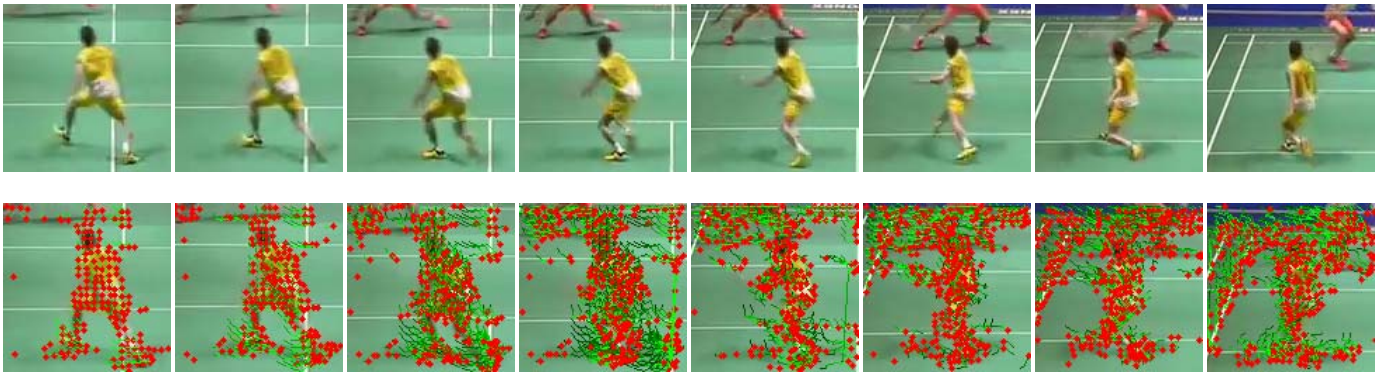

Fig. 3: Visualization of dense trajectories created for a backhand stroke. Top row: image sequence of the stroke. Bottom row: generated dense trajectories (in green) by the stroke

we used very few training samples (15 from each class) for testing, a stroke with a similar body posture for this, belonging to forehand class was not included in the training dataset. Therefore, in feature vector (histogram) the stroke was more biased towards the backhand class. A larger number of training data from various stroke styles would enabled the system to recognize the stroke.

In evaluation approach 2, two other class shots were misidentified as forehand and two backhand strokes were misidentified as forehand. While playing the other class shots, the player was initially running towards right, which generates a motion pattern similar to a forehand stroke. But there were other class strokes which exhibited the same behavior but were recognized correctly. The reason for that was the HOG feature information and the motion patterns generated in the latter part of the stroke dominated the initial motion patterns and was able to give an accurate result. But in these particular strokes, which were misidentified, the legs were stretched in a way which was very similar to a forehand stroke even in the latter part of the stroke as in Fig. 4b. Therefore, the static appearance dominated the motion patterns and the system classified it as a forehand stroke. The foremost reason for the two misidentified backhand strokes was the lack of training data. At the initial part of the strokes, the appearance and the motion of the player were biased towards backhand class. But in the latter part of the stroke, the racket hand became not visible due to occlusion and the other hand was stretched in a similar way to a forehand stroke. Latter motion of the stroke and the static appearance dominated the feature vector (histogram) and the stroke was classified as a forehand stroke. Body posture of one these strokes is shown in Fig. 4c.

We can see from the results that the number of frames belonging to each stroke does not affect the accuracy significantly as long as a major part of the stroke is included in the test frames. This means the starting frame and the ending frame do not have to be precisely marked to identify a stroke accurately. A detailed description of the variety of the number of frames belonging to each stroke class and their respective identification accuracy are listed in Table IV.

(a) (b) (c)

Fig. 4: Images of misrecognized strokes

TABLE IV: Maximum and minimum number of frames belonging to each stroke class and their respective recognition accuracy

Stroke class	Maximum number of frames belong to stroke class	Minimum number of frames belong to stroke class	Overall accuracy
Smash	33	23	100%
Forehand	32	14	97.72%
Backhand	44	15	93.34%
Other	50	21	90.90%

Overall accuracy drops slightly as the number of maximum frames that belong to a stroke class increases. However, our algorithm is able to maintain a high accuracy over ninety even when the maximum number of frames of a stroke class exceeds 50 and the difference between maximum and minimum frames span up to 29. Our method is, therefore, capable of capturing general motion patterns occurring inside the set of frames and come to an accurate conclusion.

B. Comparison with the state-of-the-art

It is interesting to compare our results with the results obtained by existing state-of-the-art methods. A detailed comparison of our work with Miyamori *et al.* [6] and Zhu *et al.* [9] is illustrated in Table V.

TABLE V: Comparison of our system with the state-of-the-art methods.

Method	Total recognition rate	Number of stroke classes
Miyamori *et al*.	97.42%	5
Zhu *et al*.	90.21%	2
Our (Evaluation approach1)	98.34%	4
Our (Evaluation approach2)	93.34%	4

The recognition rates are taken as provided in their work. It should be mentioned that in including results of Miyamori *et al*. [6], we only used their stroke recognition rate of the bottom player as we tested our work only on the bottom player. Their total recognition rate including the top player, was 91.06%. Note that our evaluation approach 1 exceeds both [6] and [9]. Miyamori *et al*. [6] exceeds our performance in evaluation approach 2. Both our evaluation approaches exceed Zhu *et al*. [9].

IV. Conclusion

In this paper we presented a novel, motion-based approach for automatic stroke recognition of badminton games. Our method is capable of giving robust and accurate results when applied to low quality and noisy broadcast videos. It also needs relatively fewer training samples to obtain results of high accuracy as shown in the results section. By training the system with a higher number of samples extracted from number of different videos would enable the system to be more robust and accurate. Although in this paper we applied our method only for badminton stroke recognition, it is applicable to any kind of sport or non-sport application which needs low-level action classification. Dense trajectory aligned HOG features are capable of capturing the general motion patterns inside a video snippet and accurately categorize it. Explicit extraction of the player, racket, or any other body part is not necessary for this method. Also prior knowledge of the domain, e.g., badminton court, is not needed. Our method is capable of recognizing strokes ranging inside a large number of frames or a relatively small number of frames with a high accuracy as long as the set of frames capture a major part of the stroke. Future work points towards automatic recognition of the spatial location of the player and the temporal location of each stroke which would enable the possibility of automatically generating a full statistical description of a sport video and automatic simulation of the scoreboard.

References

[1] Damien Connaghan, Ciarán O Conaire, Philip Kelly, and Noel E O'Connor. Recognition of tennis strokes using key postures. In *2010-21st IET conference on Irish Signals and Systems Conference (ISSC)*, University College Cork, Cork, Ireland, 2010. IET.

[2] Ewa Kijak, Guillaume Gravier, Patrick Gros, Lionel Oisel, and Frédéric Bimbot. Hmm based structuring of tennis videos using visual and audio cues. In *In proceedings of the 2003 International Conference on Multimedia and Expo, 2003. ICME'03. Proceedings*, volume 3, pages III–309, Baltimore, MD, 2003. IEEE.

[3] Ciarán Ó Conaire, Damien Connaghan, Philip Kelly, Noel E O'Connor, Mark Gaffney, and John Buckley. Combining inertial and visual sensing for human action recognition in tennis. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pages 51–56, Firenze, Italy, 2010. ACM.

[4] Terence Bloom and Andrew P Bradley. Player tracking and stroke recognition in tennis video. In *APRS Workshop on Digital Image Computing (WDIC'03)*, volume 1, pages 93–97, University of Queensland, St Lucia, Brisbane Australia, 2003.

[5] G Sudhir, John Chung-Mong Lee, and Anil K Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database.*, pages 81–90. IEEE, 1998.

[6] Hisashi Miyamori and Shun-ichi Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 320–325, Grenoble, France, 2000. IEEE.

[7] Yihong Gong, Lim Teck Sin, Chua Hock Chuan, Hongjiang Zhang, and Masao Sakauchi. Automatic parsing of tv soccer programs. In *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 167–174, Ottawa, Canada, 1995. IEEE.

[8] Gopal Sarma Pingali, Yves Jean, and Ingrid Carlbom. Real time tracking for enhanced tennis broadcasts. In *Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 260–265, Santa Barbara, CA, 1998. IEEE.

[9] Guangyu Zhu, Changsheng Xu, Qingming Huang, Wen Gao, and Liyuan Xing. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 431–440, Santa Barbara, CA, 2006. ACM.

[10] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, 1962.

[11] Shiming Xiang, Feiping Nie, and Changshui Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600–3612, 2008.

[12] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO.

[13] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.*, volume 1, pages 886–893, San Diego, CA, 2005. IEEE.

[14] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *SCIA13: Gothenburg, Sweden*, pages 363–370, 2003.