Combining Inertial and Visual Sensing for Human Action Recognition in Tennis

Ciarán Ó Conaire, Damien Connaghan, Philip Kelly and Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Dublin 9, Ireland.
{oconaire,conna}@eeng.dcu.ie
{kellyp,oconnorn}@eeng.dcu.ie

Mark Gaffney and John Buckley
Tyndall National Institute
"Lee Maltings", Dyke Parade
Cork, Ireland.
{john.buckley, mark.gaffney}@tyndall.ie

ABSTRACT

In this paper, we present a framework for both the automatic extraction of the temporal location of tennis strokes within a match and the subsequent classification of these as being either a serve, forehand or backhand. We employ the use of low-cost visual sensing and low-cost inertial sensing to achieve these aims, whereby a single modality can be used or a fusion of both classification strategies can be adopted if both modalities are available within a given capture scenario. This flexibility allows the framework to be applicable to a variety of user scenarios and hardware infrastructures. Our proposed approach is quantitatively evaluated using data captured from elite tennis players. Results point to the extremely accurate performance of the proposed approach irrespective of input modality configuration.

Categories and Subject Descriptors

I.5.4 [**PATTERN RECOGNITION**]: Computer vision, signal processing; I.4.8 [**IMAGE PROCESSING AND COMPUTER VISION**]: Sensor fusion

Keywords

Data Fusion, Activity Classification, Image Processing, Inertial Measurement Units, Accelerometers

1. INTRODUCTION

As part of a longer-term research programme, we are striving to develop a cheap, unobtrusive, near real-time and ultra-portable motion capture system that can obtain detailed and accurate 3D biomechanical information on sports player movement in large play areas such as outdoor arenas, where traditional motion capture set-ups would be expensive and problematic. Due to both the speed and explosive nature of the actions performed by high performance players, we focus on tennis as a challenging test scenario.

ARTEMIS'10 Artemis 2010

In this paper we focus on the specific objective of automatically determining the stroke (serve, forehand, backhand) played by an athlete for a given ball contact. This is an essential classification step prior to analysis of player motion dynamics, as well as providing semantic annotation to the tennis data collected by coaches as part of normal training.

A novel aspect of this work is the flexibility of the approach due to the complimentary use of low-cost visual sensing and low-cost inertial sensing, thereby making it applicable to a variety of elite or amateur sporting set-ups, and to both competitive and training scenarios. Either sensing modality may be employed for both stroke detection and classification if the required hardware is available. In addition, if both modalities are available the merging of features obtained from the two complimentary sources can be implemented and the stroke class can be determined via an early fusion methodology. This flexibility makes the use of this framework feasible for the vast majority of set-ups and scenarios. For example, in training scenarios the use of both modalities could be employed, however in match scenarios the use of inertial sensors may not be permitted and only visual sensing possible. Conversely, the inertial sensing technique provides a location independent approach that can be easily transported and adopted for use in training complexes without the required camera hardware infrastructure.

Section 2 reviews previous work in this area. A brief overview of our system is given in section 3. The proposed technique for extracting the temporal location of strokes within a match is presented in section 4. We describe the visual and accelerometer-based features we use for classification in section 5. Section 6 details our experimental setup and the high-performance tennis dataset we use for our tests. Our proposed approach is evaluated in section 7 and we discuss our findings and directions for future work in section 8.

2. RELATED WORK

There has been much research in the area of automatic recognition of human activities, with a recent review of human motion analysis given in [4]. In examining prior literature related to the work in this paper, we focus on prior work in tennis action analysis and in using accelerometers for action classification.

There has been much previous work on tennis stroke recognition. In [13], optical flow is used to extract features of

tennis motions and SVMs are used for shot classification. They classify strokes into either a left-swing or right-swing class (corresponding to backhand and forehand). Shah et al. [10] extract a skeletonization of the tennis player's body and feed an orientation histogram of this skeleton into SVN classifiers to distinguish forehand, backhand and 'neither'. Bloom and Bradley [1] detect a shot *keyframe* when the ball makes contact with the racket and use heuristics based on the player and racket locations to do stroke classification. Petkovic et al. [8] use *Pie* features and six Hidden Markov Models to classify tennis strokes as forehand, backhand, service, smash, forehand volley and backhand volley. Similar visual features to those used in this paper were used in [3] for gait analysis.

As a complementary modality to visual analysis, accelerometers are lightweight and can be worn on the body to aid the recognition process. Microphones are combined with three-axis accelerometers by Ward et al. to determine the activities of a person in maintenance and assembly tasks [12]. Dong et. al propose an activity tracking system using wearable accelerometers in [2]. In their system, accelerometers are placed on body segments and multiple models are used to determine the angles of rotation of the person's limbs. The models correspond to different intensities of motion. For example, a static model is used to determine the joint angles when the person is not moving, whereas a periodic model is used when the person is walking.

Pylvanainen describes a hand gesture recognition system using a 3D accelerometer and continuous hidden Markov models for classification [9]. HMMs are also used by Liang et al. [6] in order to produce choreographed motions for applications such as pre-production of animation, avatar control in virtual reality and game-like scenarios. Slyper and Hodgins demonstrated the use of wearable accelerometers for realtime avatar control [11].

3. OVERVIEW

In this section, we give an overview of our shot detection and classification system. We also describe the data capture infrastructure that we use as a test-bed for our experiments.

3.1 Shot analysis system

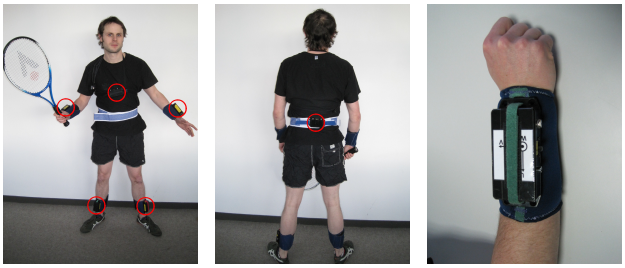
Figure 1 gives an overview of our stroke detection-and-classification system. Firstly, tennis strokes are detected. This can be done using either video or accelerometer information. This involves determining the temporal locations within a match where strokes occur. We detail our approach to stroke detection in section 4.

Each detected stroke is then classified as being either forehand, backhand or serve using pre-trained classifiers. These classifier can use data from the wearable accelerometers or from video analysis, or data from both sources.

In section 5, we describe the features that are extracted from both modalities in order to perform classification. Both sets of features are obtained from their respective sources using temporal windows located around initial detection time. We investigate the use of both SVN and K-NN classifiers for determining stroke type.

3.2 Infrastructure

In collaboration with a tennis organisation, we have instrumented an indoor tennis-court with a data-gathering infrastructure for use as a test-bed for sports and health research. This infrastructure includes overhead and baseline IP cameras with pan, tilt and zoom capability from which a tennis match can be digitally captured (see figure 3). In addition, the capture framework provides the infrastructure to capture data streams from multiple wireless accelerometers that can be placed on one or more athletes on the court. In this work we focus on using a number of custom built (±12G, 120Hz) wireless accelerometers placed onto the body of an elite tennis player. Figures 2(a) and 2(b) show the locations of the sensors on the body. Figure 2(c) shows a close-up of one of the sensors. It should be noted that these sensors are initial prototype units, and future versions will be much smaller and less obtrusive to the wearer. Inspired by the work of [11] it is intended to uses these sensors as the main input modality for full body motion capture in future work.



(a) Sensors Front (b) Sensor Back (c) Sensor

Figure 2: Accelerometer placement: (a)/(b) sensor locations on the front/back of a player; (c) Wireless inertial measurement unit (WIMU).

Figure 3: Data Capture Infrastructure: View taken from the overhead camera showing the positions of the other 8 cameras.

4. TENNIS STROKE DETECTION

Before stroke classification occurs, the temporal locations within a match where these occur must be determined. This can be determined using either video data, from an overhead camera, or wearable accelerometer data from the player's right forearm (dominant arm).

4.1 Video stroke detection

In previous work, a tennis-stroke detection approach was developed using a single overhead camera [7]. Using simple consecutive-frame-differencing, pixel blobs, corresponding to moving objects (such as the ball), were identified in

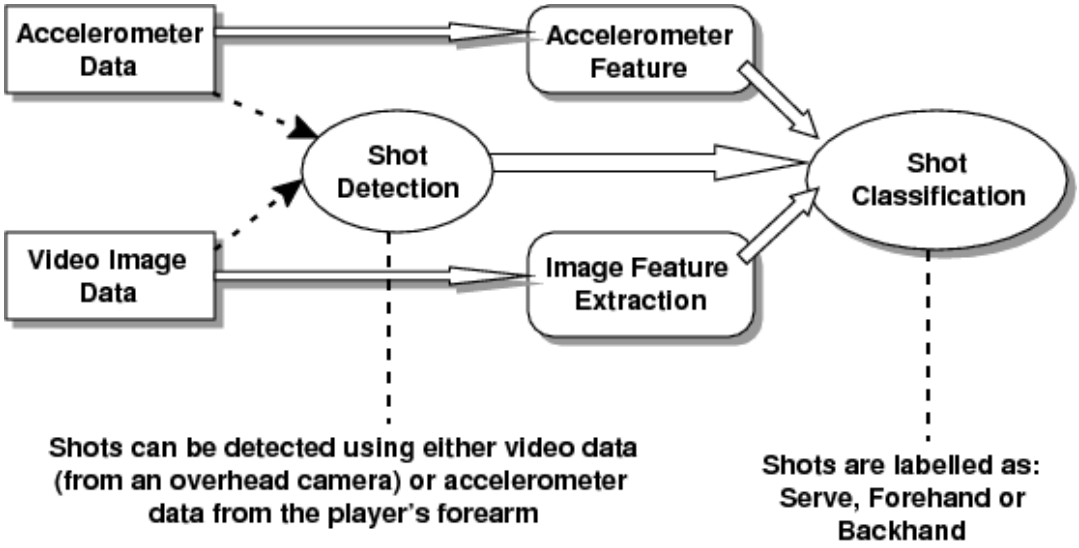


Figure 1: System overview: Our system both detects and classifies tennis strokes, using either video data or wearable accelerometer data, or data from both modalities.

each video frame. By tracking these pixel blobs in consecutive frames, non-ball pixels (caused by image noise, lighting changes and player movement) could be removed. Tracked objects that followed linear trajectories were considered to be tennis balls.

Precision and recall figures for stroke detection were 0.9429 and 0.9506 respectively, indicating that high accuracy can be achieved using this approach. However, many tennis court areas, particularly those outside, do not have an overhead view available. As such, an alternative approach based on accelerometer data is proposed here that can achieve as good as, or better, performance than that of [7].

4.2 Accelerometer stroke detection

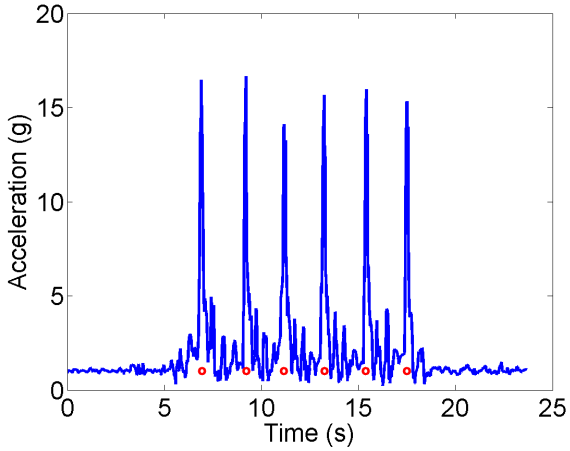


Figure 4: Shot detection using accelerometer magnitude on dominant arm: red circles show detected shots.

An accelerometer placed on a player's dominant arm will register a *spike* in its data due to the impact of the ball on the tennis racket. Detecting such data-spikes allows the automatic detection of tennis strokes, as shown in figure 4.

To detect ball contact impacts, we first compute the acceleration magnitude for each sensor sample, simply by taking the length of the 3D acceleration vector. We select the value with the largest absolute magnitude in the data. A W-second window around this peak is extracted to represent the stroke in progress. Adopting a greedy approach, this window is removed from the data and the procedure is then repeated to find the remaining strokes, until there are no acceleration magnitude peaks larger than a threshold, T. For our experiments, we used $W = 1s$ and $T = 8g$.

A window size of one second was deemed wide enough to capture the dynamics of a tennis stroke, while avoiding any additional movements performed before or after the ball impact. Threshold T was chosen based on empirical observation of the force at which a range of athletes strike the ball. Rarely does any accelerometer register readings above this magnitude unless a ball contact has been made. Subtle ball contacts, such as when the tennis ball is tapped gently over the net, are difficult to detect using this approach as the data *spike* is negligible. However, since we consider only high impact strokes (namely: serves, forehands and backhands) in this paper, we were able to achieve 100% detection accuracy.

For the remainder of the paper, we assume perfect detection accuracy of tennis strokes and focus on the classification of these strokes.

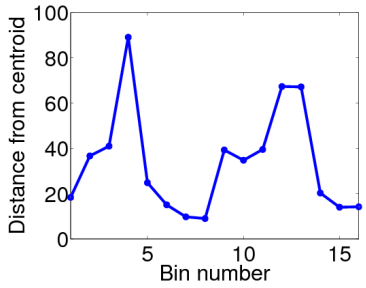
5. FEATURE EXTRACTION

In this section, we describe the features we extracted from each stroke in order to classify it into either a *serve*, *forehand* or *backhand*.

5.1 Image-based Contour Features

Figure 5 illustrates the contour features we extract from video clips of tennis strokes. In each frame, we use background subtraction to determine pixels belonging to the tennis player, as illustrated in figure 5(b). While the contour features that we use are robust to foreground holes and noise

(a) Camera image (b) Foreground (c) Radial map



(d) Contour features

Figure 5: Contour feature extraction.

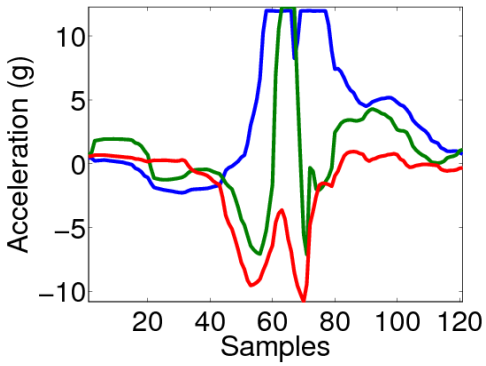


Figure 6: Accelerometer features example (a serve): the three plots correspond to the three axes (XYZ).

in the extracted silhouette, they can be adversely affected by shadows. For this reason, we use a layered background model that includes robust shadow removal[1].

We assume the player is the largest foreground connected component in the image. Using simple image morphology, we join it to other regions that are close to the player region (to account for cases where a gap in the foreground causes the arms/feet/racket to be disconnected from the player's body). To extract contour features, we divide the player foreground region into 16 *pie segments*, centred on the player centroid. For each segment, we store the largest distance of any foreground pixel from the centroid.

Over the entire stroke, we extract contour features for each video frame. We then normalise the features in order to make them invariant to the player's distance from the camera, by computing the median of all contour values and dividing all features by this value. As expected, these features are not invariant to camera viewing angle and this aspect is explored in our experiments in section 7.2.

5.2 Accelerometer-based Features

To represent a stroke in accelerometer space, we simply use the raw data. Each accelerometer sample comprises a $3D$ (xyz) vector. A single stroke is made up of a series of such samples. Example data from a serve is shown in figure 6. We normalise the stroke data by rescaling so that the variance is equal to one, in order to account for differences in stroke power.

[1]http://elm.eeng.dcu.ie/õconaire/source/

6. EXPERIMENTAL SETUP

In this section, we detail how the features we extract from each modality are used for stroke classification, as well as how we can combine both modalities to improve performance. We also describe the dataset we use to evaluate the performance of our system.

6.1 Classification

In order to classify strokes, we investigated the use of two common classifiers: Support Vector Machines (SVMs) and K-Nearest Neighbour classifiers (KNNs).

We used the SVMlight implementation of SVMs [5] with the radial basis function (RBF) kernel, as this was shown to perform well in many previous SVM studies. All feature vectors were rescaled to the range [0..1].

For the K-Nearest-Neighbour (KNN) classifier, we use a similarity measure that is robust to variations in stroke speed and power. For a given pair of tennis strokes, we compute their similarity as follows. First, we dynamically-time-warp them (minimising the sum of Euclidian distances between corresponding samples) to align them in time, thereby accounting for variations in speed. We then compare them using normalised-cross-correlation (NCC), which is invariant to stroke power.

6.2 Data fusion

If both sets of features are available, we use data fusion to improve performance. We investigated a number of different approaches to combining the two data sources for classification, including (i) simply concatenating the feature vectors and using SVMs/KNNs and (ii) learning the weight for linearly combining the similarities from both sources, then using KNNs.

While the performance was good using a fixed camera view for training and testing, performance dropped significantly when the camera view was changed. When one of the data sources is not working very well (such as when the camera viewpoint is different from the viewpoint used for training), the fusion of both data sources should account for this by (primarily) using the better data source and thereby still give good results. This was not the case in most fusion strategies we investigated.

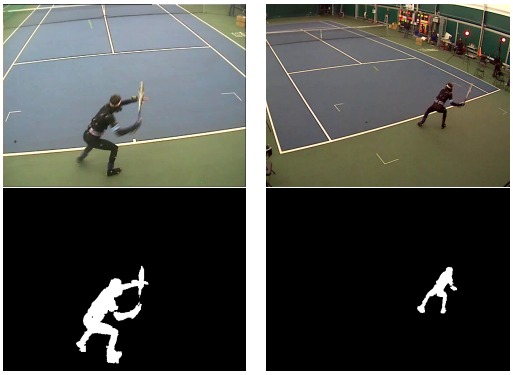
To account for the potential failure of one data source, we adopted a fusion strategy based on an *adaptive confidence weighting*, α. We define the similarity between two strokes i and j as:

$$S(i,j) = \alpha S_{acc}(i,j) + (1-\alpha) S_{vid}(i,j) \quad (1)$$

where S_{acc} is the stroke similarity in accelerometer space and S_{vid} is the similarity in the video domain. The adaptive confidence weighting α is computed from:

$$\alpha = \frac{A_{conf}}{A_{conf} + V_{conf}} \quad (2)$$

where A_{conf} is the confidence of the accelerometer data and V_{conf} is the confidence of the video data. These confidences



(a) Back view (b) Side view

Figure 7: The camera views used in our experiments, along with some illustrative detected foreground.

are computed from learned models of the expected distribution of good and bad matches. For either data source, their confidence value was computed as:

$$S_{conf} = \frac{N(x;\mu_+,\sigma_+)}{N(x;\mu_+,\sigma_+) + N(x;\mu_-,\sigma_-)} \quad (3)$$

where $S \in \{A, V\}$, x is the similarity score of the nearest neighbour in the training set, $N(x, \mu, \sigma)$ is a Gaussian model evaluated at x. $\mu+$ and $\mu-$ are computed from the training data and represent the mean similarity scores for nearest-neighbour strokes of the same-type and of difference-types, respectively. The σ values are the computed variances. Finally, KNN is used to classify strokes using the $S(i, j)$ values, which represent the combined data from both sources.

Using this approach, the weighting or *trade-off* between video and accelerometers is dynamically adapted for each new stroke to be classified, so can account for failure of one of the data sources. Our results in section 7 validate this fusion approach.

6.3 Testing Dataset

Our dataset comprises data from 5 elite, nationally ranked, tennis athletes. A total of 300 strokes were performed. Each player performed serve, forehand and backhand strokes while wearing 6 wearable wireless accelerometers. Video of the performance was captured using the cameras in the installation (figure 3). However, only 2 of the 9 cameras were suitable for classification due to occlusion and camera direction. These views are shown in figure 7.

7. RESULTS

We now detail the stroke classification accuracy that was achieved in our various experiments, using both video and accelerometer data. For video classification, we examine two different viewpoints: a back view and a side view, shown in figure 7(a) and 7(b) respectively. For accelerometer-based classification, we examine the use of all 6 wearable accelerometers, as shown in figure 2, as well as investigating how well strokes can be classified using on one accelerometer on the dominant arm (all our players were right handed). For all tests, we use a leave-one-out approach. That is, we use one player for testing and the remaining four players for training. We do this for all players and use the average classification accuracy as the performance metric. All parameters for the classifiers are learned using 4-fold cross-validation (each player's shots are in a separate fold). We used a grid search to optimise the SVM parameters and for the KNN classifier, we chose the best K from the set $\{3, 5, ..., 13, 15\}$.

7.1 Classifier comparison

Source	Classifier	Accuracy
Video (Back view)	SVM	97.02%
Video (Back view)	KNN	98.67%
Video (Side view)	SVM	89.69%
Video (Side view)	KNN	95.00%
Accelerometer (Right arm)	SVM	88.36%
Accelerometer (Right arm)	KNN	89.41%
Accelerometers (Full body)	SVM	82.43%
Accelerometers (Full body)	KNN	93.44%

Table 1: Classification results: comparing SVM and KNN classifiers shows that KNN provides greater accuracy due to its use of dynamic time warping to compare strokes and account for variations in stroke speed.

Table 1 shows the results of our tests comparing SVM and KNN classifiers using a variety of data sources. In all tests, using KNN boosts performance due to its use of dynamic time warping to account for variations in speed. In general, we found that using video gives better results than accelerometers, and using more accelerometers gives better results than using just a single accelerometer. In terms of camera viewpoint, classification accuracy is best for the back view, but that is most likely due to the fact that the player appears larger in the back view, thereby giving a more accurate foreground, rather than suggesting it to be the most useful viewing angle for classification.

7.2 Viewpoint sensitivity

Trained on	Tested on	Classifier	Accuracy
Back view	Side view	KNN	48.52%
Side view	Back view	KNN	71.10%

Table 2: Video classification view sensitivity: the visual features are sensitive to viewpoint.

While the last section suggested that video data is more useful than accelerometer data, table 2 illustrates the sensitivity of the video contour features to viewing angle changes. When the viewing angle during testing is different from the view used for training (approximately 45^o difference in this case), the accuracy is significantly affected.

7.3 Combined recognition

By combining data from both accelerometers and video, we are able to increase the classification rate further, even achieving perfect classification with all accelerometers and the camera view from behind the player (see table 3).

Our fusion strategy can also overcome the issues of viewpoint sensitivity, as shown in table 4. Despite the change in camera view, the adaptive confidence weighting can recognise that the similarity scores for the visual contours are outside the expected distribution for correct matches and

Video Data	Acc Data	Accuracy
Back view	Right Arm	98.68%
Side view	Right Arm	98.00%
Back view	Full body	100.00%
Side view	Full body	99.67%

Table 3: Results of fusion accelerometers and video data show that using both sources boosts performance (KNN classifier used)

Trained on	Tested on	ACC Data	Accuracy
Back view	Side view	Right Arm	86.73%
Side view	Back view	Right Arm	84.73%
Back view	Side view	Full body	94.08%
Side view	Back view	Full body	96.71%

Table 4: Results of fusion accelerometers (ACC) and video data show that using both sources boosts performance (KNN classifier used)

reduces the confidence in the video data, thereby relying mostly on accelerometers in difficult cases.

8. DISCUSSION

In this work, we investigated how tennis strokes can be automatically detected and classified using video-based or wearable-accelerometer-based systems in order to provide semantic information to players and coaches. Depending on the infrastructure available, only one of these type of data might be suitable and we were able to achieve high classification rates with either source. Furthermore, by combining both sources of data we were able to achieve almost perfect classification on a challenging dataset of tennis strokes performed by elite tennis athletes. Additionally, our proposed adaptive confidence fusion can robustly handle the failure of either data source and still provide high quality classification.

In future work, we want to investigate the optimal placement of accelerometers on the body for maximum classification accuracy, as well as expanding the stroke classes to include more detailed classification, such as single/double-handed backhands and ground-strokes, in order to provide a richer semantic annotation for players, coaches and tennis-enthusiasts.

9. ACKNOWLEDGMENTS

This work is supported by Science Foundation Ireland under grant 07/CE/I1147 and by the Tyndall National Institute under NAP Grant 209.

10. REFERENCES

[1] T. Bloom and P. Bradley. Player tracking and stroke recognition in tennis video. In *Proceedings of the WDIC*, pages 93–97, 2003.

[2] L. Dong, J. Wu, and X. Chen. A body activity tracking system using wearable accelerometers. In *ICME'07*, pages 1011–1014, 2007.

[3] H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletonization. In *Proc. of the Workshop on Application of Computer Vision*, October 1998.

[4] X. F. Ji and H. H. Liu. Advances in view-invariant human motion analysis: A review. *IEEE Trans on Systems, Man, and Cybernetics*, 40(1):13–24, Jan 2010.

[5] T. Joachims. Making large-scale support vector machine learning practical. pages 169–184, 1999.

[6] X. Liang, Q. Li, X. Zhang, S. Zhang, and W. Geng. Performance-driven motion choreographing with accelerometers. *Comput. Animat. Virtual Worlds*, 20(2‐3):89–99, 2009.

[7] C. Ó Conaire, P. Kelly, D. Connaghan, and N. E. O'Connor. Tennissense: A platform for extracting semantic information from multi-camera tennis data. In *International Conference on Digital Signal Processing (DSP)*, pages 1062–1067, 2009.

[8] M. Petkovic, W. Jonker, and Z. Zivkovic. Recognizing strokes in tennis videos using hidden markov models. In *VIIP*, pages 512–516, 2001.

[9] T. Pylvanainen. Accelerometer based gesture recognition using continuous hmms. In *Pattern Recognition and Image Analysis*, pages 639–646, 2005.

[10] H. Shah, P. Chokalingam, B. Paluri, S. N. Pradeep, and R. Balasubramanian. Automated stroke classification in tennis. In *ICIAR*, pages 1128–1137, 2007.

[11] R. Slyper and J. Hodgins. Action capture with accelerometers. In *2008 ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, July 2008.

[12] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1553–1567, 2006.

[13] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 431–440, New York, NY, USA, 2006. ACM.