Image and Vision Computing 32 (2014) 896–903

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Automatic annotation of tennis games: An integration of audio, vision, and learning☆

Fei Yan [a,*], Josef Kittler [a], David Windridge [a], William Christmas [a], Krystian Mikolajczyk [a], Stephen Cox [b], Qiang Huang [b]

[a] Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford GU2 7XH, United Kingdom
[b] School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom

ARTICLE INFO

*Article history:*
Received 17 September 2013
Received in revised form 7 June 2014
Accepted 4 August 2014
Available online 11 August 2014

*Keywords:*
Tennis annotation
Object tracking
Audio event classification
Sequence labelling
Structured output learning
Hidden Markov model

ABSTRACT

Fully automatic annotation of tennis game using broadcast video is a task with a great potential but with enormous challenges. In this paper we describe our approach to this task, which integrates computer vision, machine listening, and machine learning. At the low level processing, we improve upon our previously proposed state-of-the-art tennis ball tracking algorithm and employ audio signal processing techniques to detect key events and construct features for classifying the events. At high level analysis, we model event classification as a sequence labelling problem, and investigate four machine learning techniques using simulated event sequences. Finally, we evaluate our proposed approach on three real world tennis games, and discuss the interplay between audio, vision and learning. To the best of our knowledge, our system is the only one that can annotate tennis game at such a detailed level.



## 1. Introduction

The rapid growth of sports video databases demands effective and efficient tools for automatic annotation. Owing to advances in computer vision, signal processing, and machine learning, building such tools has become possible [1–3]. Such annotation systems have many potential applications, e.g., content-based video retrieval, enhanced broadcast, summarisation, object-based video encoding, and automatic analysis of player tactics, to name a few.

Much of the effort in sports video annotation has been devoted to court games such as tennis and badminton, not only due to their popularity, but also to the fact that court games have well structured rules. A court game usually involves two (or two groups of) players hitting a ball alternately. A point is awarded when the ball fails to travel over a net or lands outside a court area. The task of court game annotation then consists in following the evolution of a game in terms of a sequence of key events, such as serve, ball bouncing on the ground, player hitting the ball, and ball hitting the net.

On the other hand, building a fully automatic annotation system for broadcast tennis video is an extremely challenging task. Unlike existing commercial systems such as the Hawk-Eye [4], which uses multiple calibrated high-speed cameras, broadcast video archives recorded with a monocular camera pose great difficulties to the annotation. These difficulties include: video encoding artefacts, frame-dropping due to transmission problems, illumination changes in outdoor games, acoustic mismatch between tournaments, frequent switching between different types of shots, and special effects and banners/logos inserted by the broadcaster, to name a few. As a result of the challenges, most existing tennis applications focus only on a specific aspect of the annotation problem, e.g., ball tracking [5,6], action recognition [7]; or only annotate at a crude level, e.g., highlight detection [3], shot type classification [1]. Moreover, they are typically evaluated on small datasets with a few thousands of frames [5–7].

In this paper, we propose a comprehensive approach to automatic annotation of tennis games, by integrating computer vision, audio signal processing, and machine learning. We define the problem that our system tackles as follows:

- Input: a broadcast tennis video without any manual preprocessing and pre-filtering, that is, the video typically contains various types of shots, e.g. play, close-up, crowd, and commercial;
- Output: ball event detection: 3D (row and column of frame + frame number) coordinates of where the ball changes its motion; and ball event classification: the nature of detected ball events in terms of five distinct event labels: serve, hit, bounce, net, and null, which corresponds to erroneous event detection.

To the best of our knowledge, our system is the only one that can annotate at such a detailed level.

☆ This paper has been recommended for acceptance by M. Pantic.

* Corresponding author.
*E-mail address:* f.yan@surrey.ac.uk (F. Yan).

http://dx.doi.org/10.1016/j.imavis.2014.08.004
0262-8856/

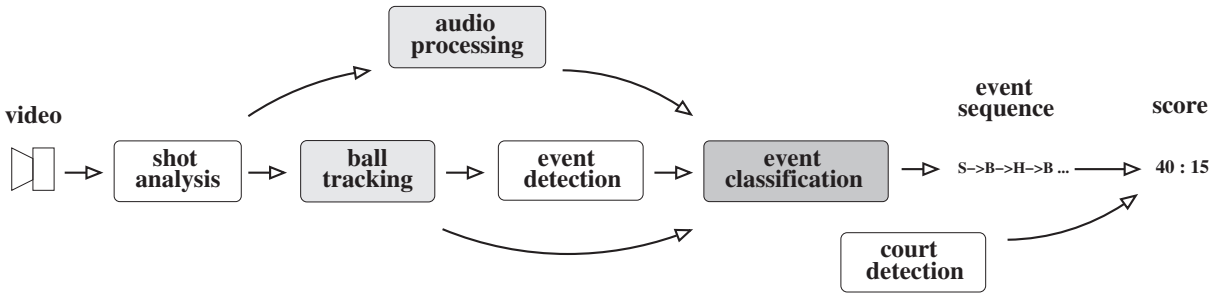


**Fig. 1.** System diagram of our proposed tennis video annotation approach. The light-shaded blocks are covers in Sections 3.1 and 3.2 respectively; the dark-shaded block is covered in Section 4.

To achieve the goal defined above, at the feature level, we improve upon our previous work and propose a ball tracking algorithm that works in the more cluttered and therefore more challenging tennis doubles games. The identified ball trajectories are used for event detection and as one feature for event classification. A second feature for classification is extracted by audio signal processing. At the learning level, we model event classification as a sequence labelling problem. We investigate four representative learning techniques and identify their advantages on simulated event sequences. Finally, our approach is evaluated on three real world broadcast tennis videos containing hundreds of thousands of frames. Discussions on the interplay between audio, vision, and learning are also provided. Note that this paper extends our preliminary work [8] by including the construction of visual and audio features, the integration of visual and audio modalities at the learning level, and a more comprehensive investigation of learning techniques.

The rest of this paper is organised as follows. Section 2 gives an overview of the proposed approach. The construction of features, including visual and audio features, is described in Section 3. Four learning techniques are then reviewed and compared on simulated event sequences in Section 4. Results on real world tennis games and discussions on the results are provided in Section 5. Finally Section 6 concludes the paper.

## 2. Overview of our approach

A diagram of our proposed system is illustrated in Fig. 1. We assume a tennis video recorded with a monocular and static camera, e.g., a broadcast tennis video. If the video is interlaced, its frames are first de-interlaced into fields, in order to alleviate the effects of temporal aliasing. For the sake of simplicity, in the remainder of this paper, we will use "frames" to refer to both frames of progressive videos and fields of interlaced videos. After de-interlacing, the geometric distortion of camera lens is corrected. De-interlacing and geometric correction are considered "pre-processing" and omitted from Fig. 1.

A broadcast tennis video is typically composed of different types of shots, such as play, close-up, crowd, and commercial. In the "shot analysis" block of Fig. 1, shot boundaries are detected using colour histogram intersection between adjacent frames. Shots are then classified into appropriate types using a combination of colour histogram mode and corner point continuity [9]. An example of the composition of a broadcast tennis video is shown in Fig. 2, where two examples of typical sequences of events in tennis are also given. The first example corresponds to a failed serve: the serve is followed by a net, then by two bounces under the net. The second example contains a short rally, producing a sequence of alternate bounces and hits.

For a play shot, the ball is tracked using a combination of computer vision and data association techniques, which we will describe in more detail in Section 3.1. By examining the tracked ball trajectories, motion discontinuity points are *detected* as "key events". Two examples of ball tracking and event detection results are shown in Fig. 3, where each key event is denoted by a red square. The detected events are then *classified* into five types: serve, bounce, hit, net, and "null", which are caused by erroneous event detection. Two features are exploited for this classification task: information extracted from ball trajectories, i.e. location, velocity and acceleration around the events (Section 3.1); and audio event likelihoods from audio processing (Section 3.2).

In addition to the features, the temporal correlations induced by tennis rules should also be exploited for classifying the events. For instance, a serve is likely to be followed by a bounce or a net, while a net almost certainly by a bounce. The focus of the "event classification" block of Fig. 1 is combining observations (features) and temporal correlations to achieve optimal classification accuracy. We model event classification as a sequence labelling problem, and provide an evaluation of several learning techniques on simulated event sequences in Section 4.

## 3. Extraction of audio and visual features

In this section, we first introduce a ball tracking algorithm which improves upon our previous work. We sacrifice completeness for conciseness, and give an outline of the complete algorithm and discuss in detail only the modifications. Interested readers are referred to [10] for details of the complete algorithm. The tracked ball trajectories are used for event detection and also as a feature for event classification. In the second half of this section, we describe the second feature for event classification that is based on audio processing.

### 3.1. Ball tracking

Ball trajectories carry rich semantic information and play a central role in court game understanding. However, tracking a ball in broadcast video is an extremely challenging task. In fact, most of the existing court game annotation systems avoid ball tracking and rely only on audio and player information [11–13,3,14]. In broadcast videos the ball can occupy as few as only 5 pixels; it can travel at very high speed and blur into the background; the ball is also subject to occlusion and sudden change of motion direction. Furthermore, motion blur, occlusion, and abrupt motion change tend to occur together when the ball is close to one of the players. Example images demonstrating the challenges in tennis ball tracking are shown in Fig. 4.

To tackle these difficulties, we improve upon a ball tracking algorithm we proposed previously [10]. The operations of the algorithm in [10] can be summarised as follows[1]:

1. The camera position is assumed fixed, and the global transformation between frames is assumed to be a homography [15]. The homography is found by: tracking corners through the sequence; applying RANSAC to the corners to find a robust estimate of the homography; and finally, applying a Levenberg–Marquardt optimiser [9,16].
2. The global motion between frames is compensated for using the estimated homography. Foreground moving blobs are found by temporal differencing of successive frames, followed by a morphological

[1] A video file "ball-tracking.avi" is submitted with this manuscript to demonstrate this algorithm.