Improved Detection of Ball Hit Events in a Tennis Game Using Multimodal Information

Qiang Huang, Stephen Cox

School of Computing Sciences
University of East Anglia, Norwich, UK

Fei Yan, Teo de Campos, David Windridge,
Josef Kittler, William Christmas

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, UK

Abstract

We describe a novel framework to detect ball hits in a tennis game by combining audio and visual information. Ball hit detection is a key step in understanding a game such as tennis, but single-mode approaches are not very successful: audio detection suffers from interfering noise and acoustic mismatch, video detection is made difficult by the small size of the ball and the complex background of the surrounding environment. Our goal in this paper is to improve detection performance by focusing on high-level information (rather than low-level features), including the detected audio events, the ball's trajectory, and inter-event timing information. Visual information supplies coarse detection of the ball-hits events. This information is used as a constraint for audio detection. In addition, useful gains in detection performance can be obtained by using and inter-ball-hit timing information, which aids prediction of the next ball hit. This method seems to be very effective in reducing the interference present in low-level features. After applying this method to a women's doubles tennis game, we obtained improvements in the F-score of about 30% (absolute) for audio detection and about 10% for video detection.

Index Terms: Scene analysis, multimodal information integration

1. Introduction

Automatic analysis of sports games is an area that is attracting considerable research attention. Such analysis has many potential applications, e.g. video retrieval of events [5, 6], object tracking [7, 2], analysis of player tactics [9] etc. Sports games videos are also attractive material for multimodal information processing research, because they contain small sets of well-defined "events" that it is possible to segment and classify [1, 4].

In this work, we focus on the detection of ball hits in a tennis game. Tennis games are comparatively rich in audio and visual information, and the event of the ball hit is the most important event in a game: its detection is key to any analysis and understanding of the game. To improve the detection of ball hits, we use a framework that integrates visual and audio information using both low-level features and high-level event-based information. Previous work [1, 8] has investigated a similar scenario, but this work used only a section of a whole game rather than long games used here. In addition, the work presented here also takes into account the impact of noise interference in the audio track on ball hit detection, which is important because the audio quality on video soundtracks is often poor. Kijak et al [4] focused more on a coarse scene segmentation rather than fine detection of events, and on processing changes of view, switching between the global view and the close-up view. This kind of visual information has limited application to ball hit detection, because changes of view are less common during a rally.

In this work, our approach is to track the trajectory of the ball in the tennis court in order to obtain useful visual information for ball hit detection. The ball trajectory usually changes abruptly after ball hit, and hence such a trajectory change can be regarded as a good indication of the hit event. Our experience is that this is a more effective technique for detection of ball hits compared with recognition of a player's action [9].

The main problem with using the visual information for ball-tracking is that it is hard to track the ball, which is a small object usually occupying only a few pixels in the image, because of variations in color, shadow, and brightness in the background of the court. In addition, when ball is near a player, it is often difficult to locate the ball accurately against the player's sportswear. Because it is necessary to consider several candidate image regions as the ball when detecting a hit, these effects can lead to "false positives".

There are several problems associated with using the audio information on the soundtrack for ball-hit detection:

1. mis-match between the overall audio characteristics of the training and the test data. To reduce the impact of this, when using our low-level audio features, we employ likelihood ratios instead of directly using the probability of each event (see Section 4).
2. extraneous sounds, such as commentators' speech, players' grunts, players' foot-scrapes etc. obscuring ball-hit sounds and causing substitution errors.
3. the strength of the ball hit sound itself is variable: volleys, for instance, cause weak sounds that are hard to detect.

To tackle these problems, our approach is to integrate audio and visual information at a high level of what we might term "events" rather than at the low level of features. We use a staged process, in which the detected visual event (in this case, a ball hit) is firstly used to coarsely determine when the ball hit occurs, and then the detected audio events refine this coarse estimate of the position of the ball hit. In addition, we also use inter-event timing information. The timing gap between any two adjacent ball hits lies within a specific range of times, and this information is helpful to predict the occurrence of next ball hit and remove some unnecessary false detections.

This paper is structured as follow: in Section 2, our theoretical framework is introduced. Detailed descriptions of visual and audio event detection are presented in section 3 and 4, respectively. Section 5 will describe how inter-timing information

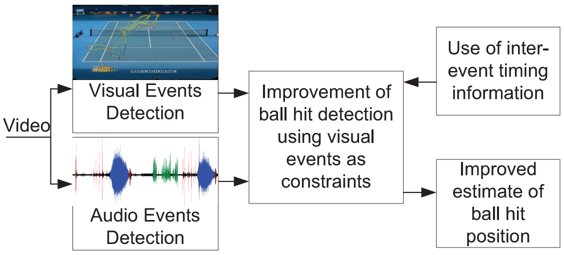
Video
Visual Events Detection
Audio Events Detection
Improvement of ball hit detection using visual events as constraints
Use of inter-event timing information
Improved estimate of ball hit position

Figure 1: Overview of our approach to ball hit detection using high-level multi modal information. Visually detected ball-hits are used as a constraint on the position of audio-detected ball hits. Inter ball hit timing information is also used.

is used to improve detection performance. The data and experimental set up is described in section 6, and section 7 analyses the results. Finally, we summarise and draw a conclusion in section 8.

2. Theoretical Framework

Figure 1 shows an overview of our approach to ball-hit detection using multimodal information. The process begins by finding the most likely sequence of visual events E_v^* together with the most likely sequence of audio events E_a^*, given the observed low-level features F and the high-level constraint of the inter-event timing information E_t. E_v^* and E_a^* can be estimated according to equation 1:

$$(E_v^*, E_a^*) = arg \max_{\{E_v, E_a\}} Pr(E_v, E_a, E_t|F) \qquad (1)$$

Equation 1 can be re-written as:

$$(E_v^*, E_a^*) = arg \max_{\{E_v, E_a\}} Pr(E_v|F)Pr(E_a|F, E_v)Pr(E_t|E_v, E_a) \qquad (2)$$

Equation 2 factors the ball hit detection into three processes:

1. ball hit detection only using visual information ($Pr(E_v|F)$)
2. ball hit detection using audio information and constraints from the detected visual events ($Pr(E_a|F, E_v)$)
3. an extra event-based constraint: inter-ball-hit timing information ($Pr(E_t|E_v, E_a)$).

The first process, visual event detection, uses ball-tracking to provide a coarse detector of ball hits, relying on the change in direction of the ball when it is struck [7]. However, using only visual information will generate false positives of ball strikes because of the reasons given in section 1. The second process uses audio information constrained by visual events. These contraints on the "window" in which the hit can take place can reduce the impact of other types of audio event and interference from noise. The third process, use of inter-ball-hit timing information predicts when the next ball hit is most likely to occur given knowledge of the time of the current hit. In addition, we find there are discrepancies between the audio and visual information in the exact positioning of a ball hit. There are two main reasons for this: firstly, the visual frame-rate is only 50 frames per second, which limits the precision at which the hit can be registered and annotated in the visual ground-truth file. Secondly, there is a delay between the racquet striking the ball and the sound being picked up by the "effects" microphone: for instance, if the microphone is 10m away, the delay is about 30 ms, which is two frames in audio processing. These two effects can conspire to produce differences in the registration of a hit of tens of milliseconds. Such discrepancies cause problems in a low-level approach audio-visual fusion approach, but are dealt with effectively in our approach.

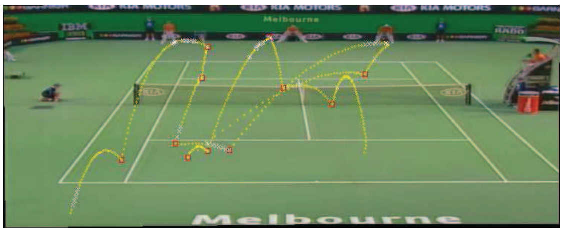

Figure 2: An example of the ball trajectory during a rally superimposed on the mosaic image of the court

3. Visual Event Detection

Our strategy to detect visually ball-hits is to track the ball position throughout the game and then hypothesise a ball hit according to the estimated ball trajectory. Below are given the most important steps in this process: for a complete description, please refer to [7].

1. Homographies between frames are computed and are used to compensate for camera motion.
2. Candidate blobs are found by temporal differencing of successive frames.
3. Blobs are then classified as ball / not ball using their size, shape and gradient direction at blob boundary.
4. "Tracklets" are established in the form of second-order (i.e. roughly parabolic) trajectories. These correspond to intervals when the ball is in free flight.
5. A graph-theoretic data association technique is used to link tracklets into complete ball tracks. Where the ball disappears off the top of the frame and reappears, the tracks are linked.
6. By analysing the ball tracks, sudden changes in velocity are detected as "ball events".

Figure 2 shows an example of the ball trajectory superimposed on the "mosaic" image of the court, which is the image of the static parts of the court during a rally. The example starts from a successful serve at the near side of the court, and ends when the ball goes out of play after six ball hits. The yellow dots in the figure indicate the positions of the candidates in the valid sets of the nodes in the shortest path, while the white crosses are interpolated ball positions. The red squares represent the detected key visual events, such as ball bounce and ball hit. Because of background interference, two ball hits on the far side of the court have been missed. From this figure, it is easy to identify the ball-hit locations and where the ball lands in the court.

4. Audio Event Detection

In previous work [3], we defined seven types of audio events for description of tennis matches, one of which was the sound of the racquet hitting the ball. Table 1 gives descriptions of each audio class and their related functions in a tennis game. For audio event detection, there are two issues to be addressed:

Table 1: Audio classes used in this work

Audio Event	Name	Function
Chair umpire's speech	UMP	Report Score
Line judge's shout	LJ	Report serve out, fault etc.
Sound of ball hit	BS	Serve, Rally
Crowd noise	CN	Applause
Beep	BP	Let
Commentators' speech	COM	
silence	SIL	-

1. distinguishing between the seven types of audio events.
2. reducing the impact of acoustic mismatches between the training and test data.

The first problem is solved in a standard maximum-likelihood framework by finding the most likely audio event given the "observed" low-level audio information, F^a , as shown in equation 3:

$$E^{a^*} = \arg\max_{E^a} \Pr(E^a|F^a) \tag{3}$$

$$\propto \arg\max_{E^a} \Pr(F^a|E^a)\Pr(E^a) \tag{4}$$

$\Pr(F^a|E^a)$ indicates a posterior probability computed using a Gaussian mixture model (GMM), and $\Pr(E^a)$ can be regarded as a prior distribution of each audio type (set equal in this paper). Equation 4 is the tranformation of equation 3 after using Bayes theorem.

To reduce the impacts of acoustic mismatch, we employ a confidence meaure (CM). The likelihood of each audio event class for a frame is estimated using the Gaussian mixture models of audio events built from the training-data, and the difference between highest log likelihood and the next highest is used as a CM for that frame. This use of a difference between likelihoods provides some immunity from mismatches between the training- and test-set channel conditions: if the mismatch is high, then all the likelihoods will be low, but the overall mismatch will be cancelled out by the differencing operation, and the differences will be relatively stable within a range. A suitable threshold for the CM corresponding to a positive detection of an audio event ball hit can be determined from the training data.

After utilizing the CM, there are still a large number of audio candidates for a ball hit E^a_{BS}. However, most of these occur outside the constraint window generated by the visual events E^v given the visual frames F^v. Equation 4 can hence be changed to:

$$E^{a^*}_{BS} = \arg\max_{E^a_{BS}} \prod_t \Pr(F^a_t|E^a_{BS})\Pr(F^v_t|E^v_{BS})\Pr(E^a_{BS}) \tag{5}$$

where F^a_t and F^v_t indicate the audio and video frame at time t. To simplify computation, $\Pr(F^v_t|E^v_{BS})$ is modelled as a normal distribution, whose mean value, t_0, is the time when each visual ball hit is detected. Because of the synchronisation problems referred to in section 2, t is allowed to vary within ±8 frames of t_0.

5. Inter-event Timing Information

Information about the time between ball-hits can be used to further constrain the number of hypothesised ball-hits. Figure 3 shows the distributions of the inter-ball-hit timing information. The two curves are an estimate of the probability distribution of the time between hits for the detected audio (red circles) and visual (blue squares) events. Both curves show a similar timing distribution, but it can be seen that the audio curve is slightly displaced to the right compared with the video distribution, which implies that the time-difference is slightly longer for audio. This is due to the synchronization problem between the visual and audio data, which has been mentioned in section 2. Since we have no means of knowing which of these two distributions is correct, we use the mean distribution as our ground-truth for the time between successive ball-hits.

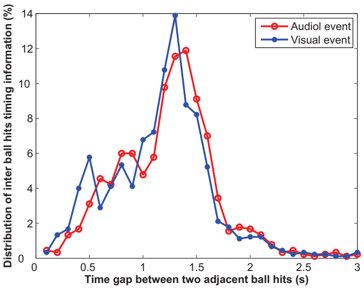

Figure 3: Distribution of the times between successive ball hits, for audio- and visually-detected events.

6. Data and Experimental Set Up

We used four video clips of a women's double game (Australia Open 2008) for training and testing. Table 2 gives some information about the length of these clips and number of ball hits in them.

Table 2: Ball events occurring in four video clips of a women's double tennis game

	VC(1)	VC(2)	VC(3)	VC(4)
Duration (mins.)	37.07	33.48	37.56	13.40
# **Ball hits**	316	250	385	135

The audio information used for training is extracted from one men's single game (Wimbledon Open 2008). The aim of using the soundtrack from a different game for training is to test the effectiveness and robustness of our method on audio event detection. When processing audio information, we segment the soundtrack into 30ms-length frames with 20ms overlapping. This means the audio frame rate is about 100Hz, higher than the visual frame rate (50 frames per second). We utilize 39-D MFCCs for each frame and build Gaussian mixture models with a variable number of components (3–8) for each audio class: we use only three components to model the ball hit class because the sound does not have too much variation.

The *F-score* is used to evaluate the detection performance. Definitions of quantities used to define this score are shown below:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$Precision = \frac{\#correctly\ detected\ ball\ hit}{\#detected\ ball\ hit} \tag{7}$$

$$Recall = \frac{\#correctly\ detected\ ball\ hit}{\#\ ball\ hit\ in\ the\ ground\ truth} \tag{8}$$

A ball-hit is considered to be “correctly detected” when the maximum likelihood value of $E^{a^*}_{BS}$ is located within the manually annotated (audio) range of a ball hit. Maximum likelihood values of $E^{a^*}_{BS}$ that are not within an audio ball-hit range are regarded as false positives, and undetected ball-hits are false negatives.

7. Results and Analysis

Table 3: Detection performance on four video clips using only audio information

Method	Metrics	VC(1)	VC(2)	VC(3)	VC(4)
Probability	*pre.*	40.86	38.24	35.20	33.04
	Rec.	62.97	69.60	68.57	54.81
	F-score	49.56	49.36	46.52	41.23
Likelihood Ratio (LR)	*Pre.*	70.29	66.34	60.49	72.55
	Rec.	75.63	81.20	77.14	82.22
	F-score	72.87	73.02	67.81	77.08
Likelihood Ratio+ Timing	*Pre.*	79.65	79.08	75.07	77.69
	Rec.	71.84	75.60	71.95	74.81
	F-score	75.54	77.30	73.47	76.23

Table 3 shows the detection performances when only audio information is used. When using the (absolute) probability of frames, detection performances are quite poor, mainly because there are many false detections caused by audio interference and acoustic mismatch. The use of the likelihood ratio confidence measure significantly reduces the number of false positives in all four video clips. After taking the inter-ball-hit timing information into account, there are improvements of the F-score values on the first three video clips, though not on the fourth video clip. We also notice that the recall values on the four video clips are lower than when only the likelihood ratio is used. This is mainly because some correct ball hits are wrongly deleted when timing information is introduced as a strong constraint.

Table 4: Detection performance on four video clips using both audio and visual information

Method	Metrics	VC(1)	VC(2)	VC(3)	VC(4)
Visual probability	*Pre.*	61.84	62.42	61.10	59.42
	Rec.	82.28	80.00	81.30	78.52
	F-score	70.61	70.12	69.76	67.64
Audio LR + Visual	*Pre.*	81.63	82.74	80.80	88.24
	Rec.	75.95	74.80	73.25	77.78
	F-score	78.69	78.57	76.84	82.68
Audio LR + Visual + Timing	*Pre.*	83.22	84.82	76.94	85.50
	Rec.	76.90	76.00	77.14	82.96
	F-score	79.93	80.17	77.04	84.21

Table 4 shows the detection performances when only visual information is used, and then when jointly applying audio and visual information to the four video clips. Using only visual information gives better performance than using only audio information, as might be expected. However, because of many false insertions, the precision values over the four video clips are still quite low. When both modalities are combined, performance is better on every metric compared with either audio or visual on their own. An analysis of the errors made by the different systems shows that this improvement comes from two aspects: better detection of volleys, and avoidance of the error of misrecognising a ball bounce as a ball hit. In a women’s double game, there are many volleys, and the audio signal they generate is often too weak to be correctly detected when only using audio information, so that it is beneficial to use visual information in this case. However, it is often difficult to distinguish ball bounces from ball hits visually, and audio information can reduce these false detections. In addition, further improvements are obtained after using the inter-ball-hit timing information.

8. Conclusion

This paper introduces our initial work on detection of ball hits by integration of high-level audio and visual information. By integrating multimodal information at the higher “event” level we minimize low-level interference, reduce computation, and solve audio/video synchronisation problems. Performance also benefits from the addition of inter-ball-hit timing information. We believe that this approach of constraining detection by using multi-modal information has general application in many audio-visual scenarios, including audio-visual speech recognition, segmentation, and understanding.

In our future work, we aim to use a similar approach in detecting the voices of the line judges and the umpires, which are also key to understanding the game. This will require improved robustness to different interferences, which we aim to achieve by integrating more context information. We also intend to extend the technique to sports games in different domains.

ACKNOWLEDGMENT: This work was supported under a UK Engineering and Physical Sciences Research Council Grant number EP/F069626/1.

9. References

[1] R. Dahyot, A. Kokaram, N. Rea, and H. Denman, “Joint audio visual retrieval for tennis broadcasts,” in *In Proceedings of ICASSP’03*, 2003, pp. 561–564.

[2] P. J. Figueroa, N. J. Leite, and R. M. L. Barros, “Tracking soccer players aiming their kinematical motion analysis,” *Computer Vision and Image Understanding*, vol. 101, pp. 122–135, 2006.

[3] Q. Huang and S. Cox, “Hierarchical language modeling for audio events detection in a sports game,” in *In Proceedings of ICASSP’10*, March 2010, pp. 2286–2289.

[4] E. Kijak, G. Gravier, L. Oisel, and P. Gros, “Audiovisual integration for tennis broadcast structuring,” in *In International Workshop on (CBMI03)*, 2003, pp. 289–312.

[5] H. Miyamori, “Automatic annotation of tennis action for content-based retrieval by integrated audio and visual information,” in *in IEEE Int. Conf. on Image and Video Retrieval*, 2003, pp. 331–341.

[6] M. Tien, Y. Wang, and C. Chou, “Event detection in tennis matches based on video data mining,” in *in IEEE Int. Conf. on Multimedia and Expo*, 2008, pp. 1477–1480.

[7] F. Yan, W. Christmas, and J. kittler, “Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1814–1830, 2008.

[8] X. Yu, C. Sim, J. Wang, and L. Cheong, “A trajectory-based ball detection and tracking algorithm in broadcast tennis video,” in *in IEEE Int. Conf. on Image Processing*, 2004, pp. 1049–1052.

[9] 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 *In Proceedings of ACM Multimeida’06*, 2006, pp. 431–440.