# JOINT AUDIO VISUAL RETRIEVAL FOR TENNIS BROADCASTS

*Rozenn Dahyot, Anil Kokaram, Niall Rea and Hugh Denman*

Electronic and Electrical Engineering Department
Trinity College, Dublin 2, IRELAND
E-mail: dahyot@mee.tcd.ie, {akokaram,oriabhan,hdenman}@tcd.ie

## ABSTRACT

In recent years, there has been increasing work in the area of content retrieval for sports. The idea is generally to extract important events or create summaries to allow personalisation of the media stream. While previous work in sports analysis has employed either the audio or video stream to achieve some goal, there is little work that explores how much can be achieved by combining the two streams. This paper combines both audio and image features to identify the key episode in tennis broadcasts. The image feature is based on image moments and is able to capture the essence of scene geometry *without recourse to 3D modelling* [1]. The audio feature uses PCA to identify the sound of the ball hitting the racket. The features are modelled as stochastic processes and the work combines the features using a likelihood approach. The results show that combining the features yields a much more robust system than using the features separately.

Keywords: Multimedia, Content Retrieval, Audiovisual analyis, PCA, Image Moments, Sports Summaries, Tennis

## 1. INTRODUCTION

Retrieval and summarisation of sports footage have received increasing interest in recent years. It is expected that the rise of home Digital media will increase the demand for easily browsable content, and hence the need for automatic content manipulation. Much of the work has concentrated on using video analysis for example, Sudhir et al. [1] extracted events from tennis using video analysis and the geometry of the court; Chang et.al [2] have used HMMs for summarizing Baseball footage. Work is emerging that considers the audio signal for spotting important events [3, 4]. In general for sports the audio signal is capable of characterising much shorter duration events than the video signal. In sports like tennis, cricket, badminton it is the short and sharp noise of the ball hitting the racket or bat that defines the basic building block action of the game. Both the audio and video signals therefore contain useful information and this work considers the use of *both* audio and video features for parsing tennis footage.

The basic unit of this game is the serve and subsequent rally or passage of play until the point is decided. Tennis summaries therefore generally contain the main court view during each of the main points. See the central frame of Figure 1 for a typical grass court shot refered as *large view* in this paper. This article presents a mechanism for extracting each rally by identifying the court view and by building a mechanism that ‘listens’ for the sound of the racket hitting the ball (referred to as a racket hit in this paper). The video sequence is first segmented into shots using the common histogram analysis technique. The task addressed by this paper is to identify each shot that is a passage of play shot containing the court view. The classification can be achieved by noting that the relevant shots contain *both* a full court view and a noise of the ball hitting the rackets. To characterise the court view we employ a recent idea that uses the implicit scene geometry [5] without recourse to 3D camera modelling. This is discussed in section 2.

Although much work has been done to characterize generic classes of sounds like speech and music, only a few authors have considered sport sound classes. These have focussed on specific sport event sound classes like bat hits in baseball [6] or dribbling, shooting, etc. sounds in basketball [3]. Previous works propose to use simple template matching techniques which do not yield satisfying results [3]. In this work the detection of racket hits is improved by using an eigenspace representation of the class of interest. A similarity measure, already used successfully in image processing [7], is computed between observations and the training eigenspace. This allows the classification of each frame of the video as a racket hit or not.

The evolution of the features is modelled as Gaussian and this admits a simple mechanism for the data fusion process as outlined in section 3. The frame level features are used to generate shot level descriptors which are used in the data fusion process. In section 4, experimental results shows the improvement in using joint visual and audio information.

## 2. VIDEO FEATURES AT FRAME LEVEL

This section presents respectively the visual and audio features used, and proposes expressions for their likelihoods.

This work was funded by EU RTN MOUMIR www.moumir.org (HP 118) and MUSE-DTV, CASMS, Enterprise Ireland.

The likelihoods express the probabilities that a frame represents a main court view and that the audio represents racket hits.

### 2.1. Frame level visual feature

The second moment of the hough transform of the edges is computed for each image [5]. This measure, noted $\mathbf{x}_v$, is used to detect frames showing a main view of the court where its value remains constant. This moment feature is low when there is strong scene geometry since the Hough space will contain compact clusters representing major lines in the image. As the large view of the court is dominated by the physical, rectangular court structure the feature works well to discriminate it without the need to resort to any 3D information (as used in [1]). Figure 2 shows the low level plateaus that correspond to the court views. The likelihood of this visual feature to be computed on a frame showing a main view of the court is then simply expressed with a gaussian law:

$$\mathcal{P}(\mathbf{x}_v|\text{Large view}) \propto \exp\left[\frac{-(\mathbf{x}_v-\mu_v)^2}{2\,\sigma_v^2}\right] \qquad (1)$$

Its mean $\mu_v$ and variance $\sigma_v^2$ are learned on training shots.

### 2.2. Frame level audio feature

Since the racket hit is a short sound between 10 to 20 ms long, we have chosen to compute the spectrogram of the audio track using a 40 ms window (duration of a frame in the video). The power spectrum of this Fourier transform, normalised by its energy, is then computed for each window and corresponds to our audio features $\mathbf{x}_a$.

**Eigenspace representation.** $K$ audio features corresponding to racket hits are collected. A Principal Component Analysis (PCA) is then performed over this training database. $J$ eigenvectors corresponding to the $J$ highest eigenvalues are retained to span the eigenspace $F$.

**Distance from the feature space.** A common way to measure the similarity of an unknown observation $\mathbf{x}_a$ with the training cloud, is to compute the distance between $\mathbf{x}_a$ and the eigenspace $F$. This *Distance From Feature Space* (DFFS) is defined as [7]:

$$\text{dffs}(\mathbf{x}_a) = \|\mathbf{x}_a - \mu_a - \text{U}^{\text{T}}(\mathbf{x}_a - \mu_a)\| \qquad (2)$$

where $\mu_a$ is the mean of the audio features, and U is the matrix collecting the $J$ eigenvectors computed in the learning step with PCA.

**Likelihood of having a Racket hit.** Assuming a uniform distribution over the eigenspace $F$, the likelihood of having a racket hit can be approximated [7, 8] using the likelihood of the reconstruction error :

$$\mathcal{P}(\mathbf{x}_a|\text{Racket hit}) \propto \exp\left[\frac{-(\text{dffs}(\mathbf{x}_a))^2}{2\,\sigma_a^2}\right] \qquad (3)$$

The variance $\sigma_a^2$ is estimated using the mean value of the eigenvalues $\{\lambda_j\}_{j>J}$ in $F^{\perp}$ [7].

## 3. VIDEO FEATURES AT SHOT LEVEL

The frame level video features are processed to generate shot level features. These features allow access to higher level content information, in classifying shots as rallies $R$ or not $\overline{R}$.

**Shot level visual feature.** The feature $\mathbf{x}_v^s$ considered to represent the visual content of the shots corresponds to the mean of the error square computed over the shot:

$$\mathbf{x}_v^s = \mathbb{E}_{\mathbf{x}_v\in s}\left[\left(\frac{\mathbf{x}_v-\mu_v}{\sigma_v}\right)^2\right]$$

Using the mean allows the shot level visual feature to not be dependent of the duration of the shot. Then, the likelihood is simply expressed using a gaussian law:

$$\mathcal{P}(\mathbf{x}_v^s|s=R) \propto \exp-\left[\frac{\mathbf{x}_v^s}{2}\right] \qquad (4)$$

**Shot level audio feature.** The feature $\mathbf{x}_a^s$ considered to represent the audio content of the shots corresponds to the minimum of the similarity measure DFFS computed over the shot:

$$\mathbf{x}_a^s = \min_{\mathbf{x}_a\in s}\left\{\frac{\text{dffs}(\mathbf{x}_a)}{\sigma_a}\right\}$$

The likelihood of the audio features of a shot to be a Rally is modelled as :

$$\mathcal{P}(\mathbf{x}_a^s|s=R) \propto \exp-\left[\frac{\mathbf{x}_a^s}{2}\right] \qquad (5)$$

**Fusion of audio and visual information.** Assuming the independence of audio and visual data, the likelihood using both audio and visual features has simply been computed by:

$$\mathcal{P}(\mathbf{x}_a^s,\mathbf{x}_v^s|s=R) = \mathcal{P}(\mathbf{x}_a^s|s=R)\times\mathcal{P}(\mathbf{x}_v^s|s=R) \qquad (6)$$

## 4. EXPERIMENTAL RESULTS

The following results have been computed over 2949 frames of a video sequence of an outdoor tennis match (Pierce Vs Serna). This contains 18 shots including 5 rallies. Figure 1 shows a selected frame from three different shots (crowd, close view of the player, large view of the court), and their corresponding audio features (or spectrogram) computed over the shots. Note how distinctive the racket sounds are as compared to crowd cheering or speech (observed in the close view shot).

**Training features.** The training has been performed us-

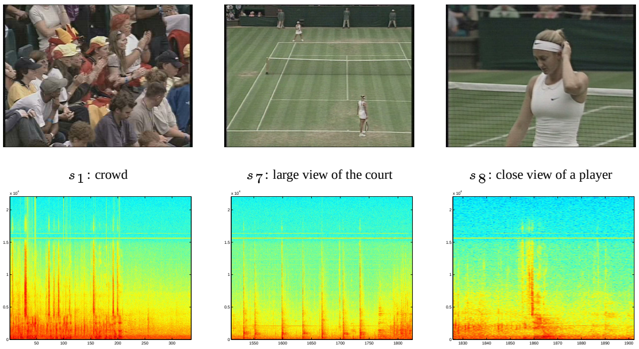

$s_1$ : crowd
$s_7$ : large view of the court
$s_8$ : close view of a player

**Fig. 1**. Frames extracted from 3 different shots $\{s_1, s_7, s_8\}$ and their corresponding spectrogram computed over the audio track of the shots.

ing a sequence of an other match. The dimension of the eigenspace has been chosen as $J = 10$ corresponding to 75% of the information over the training data [7].

**Audio and visual features.** Figure 2 presents the ground truth over the sequence: shots $\{5, 7, 11, 14, 17\}$ belong to our class of interest $R$. The second curve shows the evolution of the second moment of the hough transform : some constant plateaus appear for shots with large view of the court, but also sometimes for close view of the players. This implies that some false alarms for detecting rally shot may appear when using only visual features. The last curve presents the similarity (DFFS) computed over the sequence. Low values indicate high probabilities to have racket hits.

Figure 3 shows the receiver operating characteristic (ROC) curve computed with our racket hit detector using several tennis audio test tracks: setting a threshold at $0.3$ allows to detect more than 90% of the racket hits for less than 1.5% of false alarms. In the Pierce sequence, 19 of the 20 racket hits are detected without any false alarms. This prelimary result on sport event sound detection shows that simple learning techniques can be successfully used with better results than the template matching technique [3].

**Rally Shot Detection.** Figure 4 presents the log-likelihoods of the 18 shots of the test using respectively from top to bottom, the logarithm of the expressions 4, 5 and 6. Red crosses indicates the rally that are to be detected. Blue crosses represent the other kinds of shots. The arrows in figure 4 highlight the two highest potential false alarm shots using visual information. These represent the same camera view of one player. Figure 5 presents two frames extracted from those shots. To compare the results between the three methods, the highest log-likelihood of a non-rally shot is computed:

$$\alpha_{fa} = \max_{s \in fa} \log \mathcal{P}(\mathbf{x}_a^s, \mathbf{x}_v^s | s = R)$$

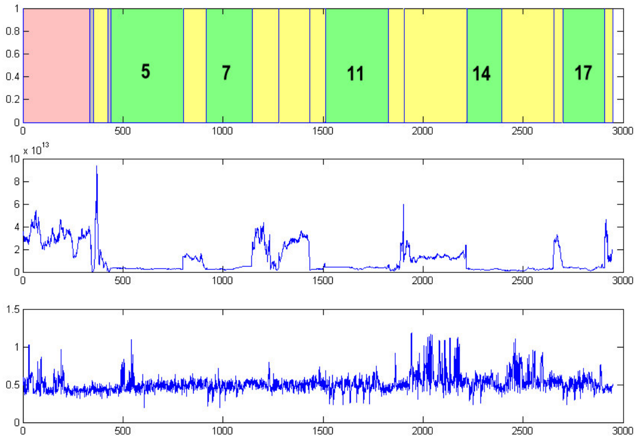


**Fig. 2**. From top to bottom : ground truth over the sequence (rose: crowd; green: large view of the court, yellow: close view of the players, grey: dissolve; black: cut), moment of the hough transform (visual features), and the dffs computed using the audio features.

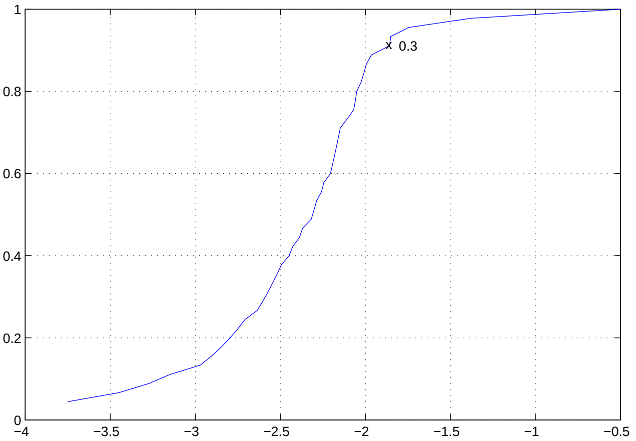


**Fig. 3**. Racket hit detection: the detection rate w.r.t. the false alarm rate (Log) using the DFFS (ROC curve).

where $fa$ collects all the non rally shots (potential false alarms). The lowest log-likelihood of a rally shot is also computed:

$$\alpha_{md} = \max_{s \in md} \log \mathcal{P}(\mathbf{x}_a^s, \mathbf{x}_v^s | s = R)$$

where $md$ collects all the rally shots (potential missed detections). Then the following measure can be used to assess the accuracy and efficiency in setting a threshold for classification of the shots:

$$\Delta = \alpha_{md} - \alpha_{fa}$$

Table 1 presents $\Delta$ for the three tests (cf. fig. 4). The negative value obtained using only the visual information indi-

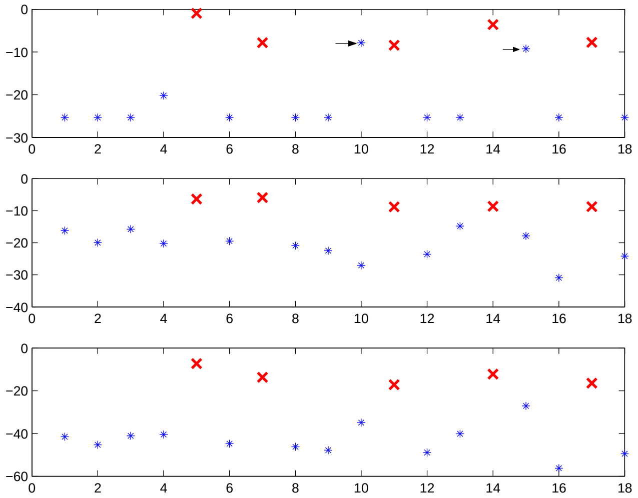
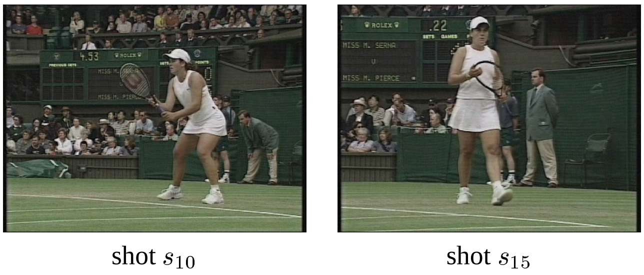


**Fig. 4**. From top to bottom: log-likelihood of the shots using only visual features, only audio features, and finally both features.


shot $s_{10}$

shot $s_{15}$

**Fig. 5**. Extracted frames from potential false alarms shots using visual information alone.

cates that it is impossible to fix a threshold allowing $100\%$ of good detection for $0\%$ of false alarms. This is possible using only audio features or using jointly audio and visual ones. But the interval to fix the threshold is bigger in using jointly audio and visual information which implies a better discrimination between the class of rally-shots and its complementary $\overline{R}$. In this experiment, visual features used alone do not allow to classify shots in the class of interest $R$ without false alarms. On the contrary, both audio features used alone and jointly with visual ones allow this. There is an advantage in joint audio visual analysis since the difference in likelihood between the two classes $R$ and $\overline{R}$ is bigger.

| | Visual | Audio | Visual & Audio |
|---|---|---|---|
| $\Delta$ | -0.5590 | 6.0106 | 9.8897 |

**Table 1**. Difference of the log-likelihoods of the first potential missed detected and the first potential false alarm.

## 5. CONCLUSION

This paper has presented a new scheme for the combination of audio and video features for extracting tennis rallies from broadcast footage. The results show that there is a substantial improvement in classification when both types of features are combined. Of particular interest is the low computational cost of the methods for feature extraction and data fusion. It is interesting to note that for broadcast sports footage the audio and video quality is very high, because it is a principal revenue earner for the broadcaster. Also during events like Tennis, there is little crowd noise during the main playing activity. This implies that there is a high likelihood of success for audio analysis of the key game events, reflected by the results shown here. Note that our current audio features are normalised which implies the loss of volume information. Although our results show that the current features work well, future work will incorporate this relevant energy feature. Our current work revolves around applying similar ideas to cricket and other ‘fixed view’ events.

## 6. REFERENCES

[1] F. Sudhir, J.C.M. Lee, and A.K. Jain, “Automatic classification of tennis video for high-level content-based retrieval,” in *proceedings of International workshop on Content-Based Access of Image and Video Databases (CAIVD’98)*, 1998.

[2] P. Chang, M. Han, and Y. Gong, “Extract highlights from baseball game video with hidden markov models,” in *proceedings of International Conference on Image Processing (ICIP)*, Rochester,NY, September 2002.

[3] D. Zhang and D. Ellis, “Detecting sound events in basketball video archive,” Tech. Rep., Electrical Engineering Department of Columbia university, 2001.

[4] Erwin M. Bakker and Michael S. Lew, “Semantic video retrieval using audio analysis,” in *proceedings of International Conference on Image and Video Retrieval (CIVR)*, London,UK, July 2002, pp. 271–277.

[5] H. Denman, N. Rea, and A. Kokaram, “Content based analysis for video from snooker broadcasts,” in *proceedings of International Conference on Image and Video Retrieval (CIVR)*, London,UK, July 2002.

[6] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for tv baseball programs,” in *proceedings of ACM Multimedia Conference*, Los Angeles, California, October 2000.

[7] B. Moghaddam and A. Pentland, “Probabilistic visual learning for object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, Juillet 1997.

[8] D.J.C. MacKay, “Probable network and plausible predictions - a review of practical bayesian methods for supervised neural networks,” *Network*, pp. 469–505, 1995.