# MULTI-MODAL AUDIO-VISUAL EVENT RECOGNITION FOR FOOTBALL ANALYSIS

Mark Barnard Jean-Marc Odobez
Samy Bengio*
Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O. Box 592, CH-1920 Martigny, Switzerland.
{barnard, odobez, bengio}@idiap.ch

**Abstract. The recognition of events within multi-modal data is a challenging problem. In this paper we focus on the recognition of events by using both audio and video data. We investigate the use of data fusion techniques in order to recognise these sequences within the framework of Hidden Markov Models (HMM) used to model audio and video data sequences. Specifically we look at the recognition of *play* and *break* sequences in football and the segmentation of football games based on these two events. Recognising relatively simple semantic events such as this is an important step towards full automatic indexing of such video material. These experiments were done using approximately 3 hours of data from two games of the Euro96 competition. We propose that modelling the audio and video streams separately for each sequence and fusing the decisions from each stream should yield an accurate and robust method of segmenting multi-modal data.**

## INTRODUCTION

With the rapid growth in the amount of multi-modal data being generated there is a need for reliable systems to automatically annotate such data. In this paper we focus on the recognition of events by using both audio and video data. Specifically we look at the recognition of *play* and *break* sequences in football and the segmentation of football games based on these two events. *Play* is defined as the ball being in normal play and *break* is when play has

*The authors acknowledge financial support provided by the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The NCCR are managed by the Swiss National Science Foundation on behalf of the Federal Authorities. This work has been performed partially with in the frameworks of the "Automatic Segmentation and Semantic Annotation of Sports Videos (ASSAVID)" project and the "Learning for Adaptable Visual Assistants (LAVA)" granted by the European IST Programme.

ceased for some reason such as, a foul, the ball going out of the field or a goal.

The segmentation of football into *play* and *break* sequences is an important task. Given the huge amount of video material current being generated manually indexing this material is prohibitively time consuming and expensive. Therefore it is important to develop an accurate and efficient technique for automatically indexing this material. In the data we have used *break* constituted 45 percent of the total time, so a segmentation into *play* and *break* provides a significant information reduction. It should be noted that in our approach to the problem of segmenting *play* and *break*, we have not based the segmentation on shot boundaries. This is important because *play* and *break* are semantic classifications that do not always adhere to shot boundaries. It is often the case that a *play* or *break* sequence will run over a number of shots and, more importantly, it is sometimes the case that a single shot will contain both *play* and *break* sequences.

The video data we are concerned with here is composed of two streams, audio and video. While some work has been done on the recognition of events within video material, this has usually focused on using either the audio or video stream in isolation. Some work has been done on the classification of television broadcast genres using the audio stream alone [5] [11]. However work in this area has concentrated on classification using the video stream. Peng Xu *et al* [15] have proposed a rule based system using video information for *play/break* segmentation of football. This work was extended to use *Hidden Markov Models* (HMMs) to model the *play* and *break* sequences and a dynamic programming algorithm to perform the segmentation [13]. HMMs have also been trained using video motion information in order to recognise events in basketball [14]. HMMs have been used with audio and video features in a scene classification task [7] and a video shot segmentation task [2]. A good review of techniques for the analysis of multi-modal data is provided by Wang, Liu and Huang [16].

In our approach we introduce the use of data fusion techniques into an HMM event recognition framework. Based on results of using multi-modal features in other fields, such as audio-visual speech recognition [6], we believe the fusion of multiple streams of data will improve both the accuracy and the robustness of the system. We will investigate the use of data fusion by low level feature vector concatentation, *early fusion*, and also by the high level combination of the decisions from each data stream, *late fusion*. In this case we use audio and video features as the data streams. In the next section we discuss the audio and video features to be used in our experiments. Next we introduce the methods we used for modelling multi-modal sequences. We then present the results of experiments comparing the performance of these various methods on the same data set.

## AUDIO AND VIDEO FEATURES

A low level set of audio and video features were selected to be used in these experiments. These low level features were selected so as to demonstrate the generality of the technique we propose to use. This differs from the approach of developing a higher level set of features specifically for the task of event recognition in football games.

The visual features $X_v^t$ at time $t$ are based on motion, and were used in this experiment to characterise the dominant motion model over the entire image field of view. More precisely, let $d_\Theta(p)$ denotes the displacement at position $p \in \mathcal{R}$ between two consecutive images $I_t$ and $I_{t+1}$. $\Theta$ denotes the parameters of the motion model, in this case an affine model, and $\mathcal{R}$ denotes the set of valid (real valued) image coordinates. The parameters of the dominant motion are first estimated using a robust estimator [10] that allows for outliers in the data. This estimation leads to the definition of a support region $\mathcal{R}_{\hat{\Theta}}$ that contains the image points that agree with the dominant motion, usually the background pixels. It is given by :

$$\mathcal{R}_{\hat{\Theta}} = \{p_1 \in \mathcal{R}/p_2 = p_1 + d_{\hat{\Theta}}(p_1) \in \mathcal{R} \text{ and } |I_{t+1}(p_2) - I_t(p_1)| < Thresh\} \quad (1)$$

The first motion measure $X_v^t(1) = d_\zeta$ characterises how well the estimated global motion model, which usually captures the image displacements that are due to the camera motion (panning, zooming etc), can actually model the displacement of points between two consecutive frames. It is defined as the ratio $\frac{|\mathcal{R}_{\hat{\Theta}}|}{|\mathcal{R}|}$, where $|\cdot|$ denotes cardinality. The second measure corresponds to the average of the motion amplitude, computed using the estimated motion model and over the entire image field of view, that is :

$$X_v^t(2) = \frac{1}{|\mathcal{R}|} \sum_{p \in \mathcal{R}} \|d_{\hat{\Theta}}(p)\|$$

The third feature is a ratio of the likelihood of no background motion and the likelihood of background motion, and can be shown to be given by [3] :

$$X_v^t(3) \propto \ln\left(\frac{\sigma^2_{\hat{\Theta}}}{\sigma^2}\right) \text{ with } \sigma^2_{\hat{\Theta}} = Var(I_{t+1}(p + d_{\hat{\Theta}}(p)) - I_t(p), p \in \mathcal{R}_{\hat{\Theta}}) \quad (2)$$

and $\sigma^2 = Var(I_{t+1}(p) - I_t(p), p \in \mathcal{R}_{\hat{\Theta}})$. These video features were extracted at the standard PAL video frame rate of one frame every 40ms.

The audio signal extracted from the broadcast tapes contained only sounds associated with the football game, such as the crowd cheering, the referee's whistle and the sound of the ball being kicked. In order to characterise this audio stream, 12 LPC Cepstral coefficients with the log energy, delta and acceleration coefficients were extracted from the raw audio signal. These are a set of robust audio features commonly used in speech recognition and in other audio recognition tasks [12], delta being the first temporal derivative of the signal and the acceleration being the second derivative. These features were included in order to characterise the dynamics of the signal. The audio features were extracted every 10 ms using a window size of 25 ms.

This produces two streams of data, $X_v$ the video stream and $X_a$ the audio stream. We have sampled them at the standard sampling rates for each mode, audio at 100 times per second and video at 25 times per second.

## MULTI-MODAL SEQUENCE RECOGNITION

The most common method currently used to model sequences of data are *Hidden Markov Models* (HMMs) [12]. HMMs are a statistical method of modeling temporal relations in sequences of data. The data is characterised as a parametric stochastic process and the parameters of this process are automatically estimated from the data. The data sequence is factorised over time by a number of hidden states $N$ and emissions from these states. The emission from each state is probabilistic and depends only on the current state. HMM training can be carried out using the *Expectation-Maximisation* (EM) algorithm and sequence decoding and recognition using the Viterbi algorithm [12]. When used in classification tasks a separate HMM is trained for each class to be recognised. So if we have $m$ classes $(k_1, \ldots, k_m)$ and data $X$ then during recognition the classification is given by finding the model $M$ that maximises the probability of the model given the data $P(M|X)$. So the selected class is

$$k^* = \arg\max_k P(M_k|X). \tag{3}$$

Using Bayes rule and assuming an equal prior on the class we get

$$k^* = \arg\max_k p(X|M_k). \tag{4}$$

The fusion of redundant information from different sources can reduce overall uncertainty and increase the accuracy of a classification system. Fusion can take place at different stages in the recognition process. In *early fusion* techniques the data is combined and then recognition is performed on this combined data. The most common method of *early fusion* is to concatenate the feature vectors from the different modes. This technique involves aligning and synchronising the data so as to form one combined data stream. In the case of audio and video streams, the audio data $X_a$, and the video data $X_v$ are concatenated to form a single audio-video data stream $X_{av}$. A single HMM is then trained for each class using this concatenated stream. Given that audio and video are usually sampled at different rates, this involves subsampling or oversampling one of the streams in order to synchronise them. In this case the selected class is

$$k^* = \arg\max_k p(X_{av}|M_k). \tag{5}$$

This *early fusion* approach, however, does not allow for asynchronicity and differences in temporal structure between the different modalities.

One solution when this assumption of state synchronicity cannot be made for the data is the use of a *late fusion* technique in which separate HMMs are independently trained for each class using the data from each stream of data. So if we have $J$ streams of data and $M$ classes the number of HMMs is $J \times M$. The decisions from each of these independent HMM classifiers is then combined to produce a classification of the sequence. In this *late fusion* technique, decisions take the form of some sort of score or classification of each stream, for example a posterior probability or log likelihood. One way of combining these decisions when they represent likelihoods and are assumed to be independent given the model is by using the product rule

$$k^* = \arg\max_k \prod_{j=1}^{J} p(X_j|M_k). \tag{6}$$

A comprehensive review of methods for combining classifiers is provided by Kittler et al [8].

In order to implement this *late fusion* approach we model the audio and video separately and then combine the likelihoods from each stream. We also introduce a weighting factor $\omega$ on the likelihoods from each stream. The likelihood outputs from the audio model and the video model are combined according to:

$$p(X|M_k) = p(X_a|M_{ak})^{\omega}.p(X_v|M_{vk})^{(1-\omega)}, \tag{7}$$

where $p(X_a|M_{ak})$ is the likelihood of the audio stream given the audio model, $p(X_v|M_{vk})$ is the likelihood of the video stream given the video model and $\omega$ is the weighting factor on the streams.

## EXPERIMENTS

The data used in these experiments was provided by the BBC sports library under the European Union Information Society Technology (EU IST) project ASSAVID. This data consists of approximately 171 minutes of football from the Euro96 competition: approximately 94.30 minutes of *play* and 76.61 minutes of *break*. This was made up of two games, the first game England vs Switzerland and the second game Italy vs Czech Republic. As was noted in the introduction the data was labelled on a semantic basis and not on the basis of shots and shot boundaries. The length of *play* and *break* sequences was extremely variable. The *play* sequences had a mean length of 19.53 seconds with a variance of 302.73 and the break sequences had a mean length of 14.27 seconds with a variance of 175.28.

### Sequence Recognition Experiment

The first experiment conducted was the recognition of sequences of *play* and *break* that had been segmented by hand. The total number of *play* sequences was 285 with 134 for training, 51 for validation and 100 for testing. In addition to this, 320 *break* sequences were segmented with 154 for training, 66 for validation and 100 for testing. Fully connected (ergodic) HMMs were used in these experiments and the observation in each state was modeled by a Gaussian mixture model. Models were trained using the audio stream only and the video stream only and also, to implement the *early fusion* approach, the audio and video features vectors were concatenated and used to train models. To concatenate the two streams the video was oversampled by a factor of four. The *late fusion* method was implemented by combining independently modelled audio and video streams. This combination was done using Equation 7. The optimal value for the weighting factor $\omega$ was determined by selecting the value that gave the highest average *log likelihood* on the validation set.

In order to find the optimal number of states and Gaussians for each data stream model, a number of different combinations of states and Gaussian were tested using the training and validation data. The optimal number of states and Gaussians for the HMMs was selected by finding the model trained by EM on the training data that produced the highest average *log likelihood* on the set of validation sequences. For *play* these were, 20 states and 15 Gaussians per state for audio, 14 states and 15 Gaussians per state for video and 13 states and 15 Gaussians per state for concatenated audio-video. For *break* these were, 20 states and 15 Gaussians per state for audio, 19 states and 5 Gaussians per state for video and 7 states and 5 Gaussians per state for concatenated audio-video. The performance of the models was measured in terms of three different errors: the false acceptance rate (FAR) which is the percentage of *play* recognised as *break*; the false rejection rate (FRR) which is the percentage of *break* recognised as *play*and the half total error rate (HTER) which is the mean of the FAR and FRR.

The decision was taken by applying the log likelihood ratio criterion: if

$$\log p(X|M = play) - \log p(X|M = break) > \Delta \quad (8)$$

then it is *play*. The value of $\Delta$ is chosen on the validation set in order to obtain the Equal Error Rate (FAR = FRR).

The relationship between the FAR and the FRR can be seen by plotting both errors as a *Detection Error Tradeoff* (DET) curve [9]. This type of curve clearly shows the trade off between false rejection and false acceptance rate. The threshold used in recognition tests was the threshold at the EER point on the DET curves generated from the validation set using the models selected with the optimal topology. The set of these curves for the audio, video, audio-video models and the fusion of audio and video is shown in Figure 1.

The optimal model for each mode was then applied to the set of test sequences. Table 1 shows the results on the test set using the threshold that produced an EER on the validation set. From these results the advantage of using both audio and video data for the sequence recognition task is clear.

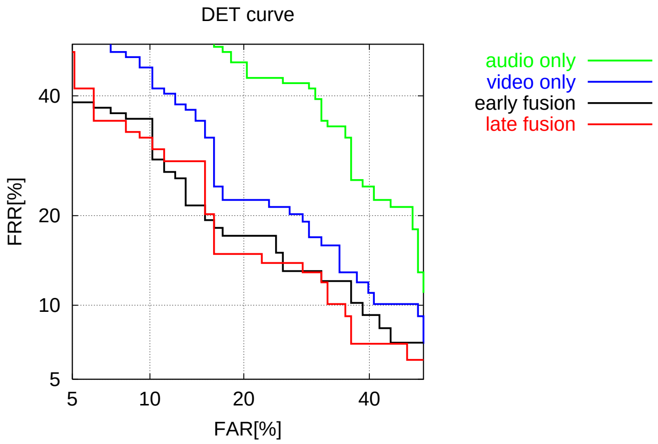


Figure 1: DET plot for validation set. This shows the performance of each modelling technique, with the false rejection rate plotted again the false acceptance rate.

Also the use of *late fusion* by combining the decision from each stream provides an improvement over *early fusion* by feature vector concatenation.

### Sequence Segmentation Experiment

In the next experiment an unsegmented piece of football data was automatically segmented into *play* and *break* sequences. The data was divided into four sections: the first and second half of both games. Models were trained on the pre-segmented *play* and *break* sequences from each of the four data sections in turn and then tested on the other three sections. This will give an indication of the ability of the HMMs to generalise both within one game and also between games. The sequences were sampled at each second with a sliding window of three seconds. This window is much shorter than the average length of the sequences. However given the large variance of the sequence lengths in the training set and the use of fully connected HMMs this should not have too much effect on the results. So for each 3 second window in the section of data we are segmenting we produce a likelihood of *play* and a likelihood of *break*.

In order to segment one half of a football game we need some way of modelling the long term structure of the game. In this case we used a 2 state fully connected HMM to model the transitions between the *play* events and the *break* events. The transition probabilities for this HMM were determined by counting the number of transitions in the section of the game used for training. The emission from each state of this HMM is given by the likelihood of *play* and *break* computed from the 3 second data window centered at each second in the section of the game we are segmenting. This HMM was then decoded for each section of the game using the Viterbi algorithm [12]. We measure the accuracy of the segmentation by comparing the classification

| Mode | FAR | FRR | HTER |
|---|---|---|---|
| Audio only | 30.6 | 40.4 | 35.5 |
| Video only | 19.3 | 22.2 | 20.8 |
| Early fusion | 20.4 | 17.3 | 18.8 |
| Late fusion | 16.3 | 15.1 | 15.7 |

Table 1: Results for each of the modelling techniques on the test set. These are results for the two-class problem of classifying *play* vs *break* in football data. The results use the *a priori* EER threshold taken from the validation set. For a random classifier the values of FAR, FRR and HTER would all be 50.

given by the Viterbi decoding at each second to the labeling of the data for that second.

The results for training on each section of data in turn and testing on the other three sections of data using the *late fusion* technique are shown in Table 2. Table 3 shows a summary of the results for the different methods used in these experiments. This shows that while using motion features alone produces good results this can be improved by the addition of the audio stream using the *late fusion* method.

While there is an increase in accuracy, the key contribution of the audio stream is an increase in robustness. This can be seen in last two columns of Table 3. The audio recognition rate is almost constant over all the test sets regardless of whether they are from the same game as the training set or not. The motion however performs noticably worse when the test set is from a different game. This lack of robustness to changes in game is even more pronounced in the results of the *early fusion* technique. By using the *late fusion* method we can significantly improve the robustness of the system to changes in game.

## CONCLUSION

In this paper we have proposed the use of both audio and video features to recognise events in football. In our approach we model the audio and video streams separately using HMMs. We then use *late fusion* to combine the decisions of the audio and video streams to form a single recognition decision. In order to test the effectivness of this method we compared it to modelling each stream alone and also the two streams combined using *early fusion* through concatenation of the feature vector. It can be seen in the results that the *late fusion* technique provides the most accurate recognition of sequences. This technique also provides the most accurate segmentation of football into *play* and *break* sequences. The paired Students t-test was used to test whether the improvement in recognition rate produced by the addition of the audio data is statistically significant. This test was performed on the results from using motion only and the results from using audio and video late fusion over the entire test set. It showed that the improvement

| Test sets | Training sets | | | |
|---|---|---|---|---|
| | Game 1 1st half | Game 1 2nd half | Game 2 1st half | Game 2 2nd half |
| Game 1 1st half | 84.5 | 83.2 | 80.6 | 82.3 |
| Game 1 2nd half | 85.5 | 87.9 | 79.2 | 80.3 |
| Game 2 1st half | 88.4 | 87.3 | 90.7 | 87.6 |
| Game 2 2nd half | 87.5 | 85.4 | 86.5 | 88.6 |

Table 2: The percentage recognition rates for the segmentation of football tapes using *late fusion* by combining the decision from each stream. The recognition rate for each tapes is shown when tested with the models trained on each of the other tapes. Note the diagonal shows the training performance.

is statistically significant with the probability of the null hypothesis being 0.003.

This shows the ability of statistical models such as HMMs to model sequences of data given simple low level features. It also highlights the advantage of being able to model each stream of data using the optimal model for that stream and then combining the decisions from the models to classify a sequence. We feel that these results could be improved further by improving the motion features and also by the introduction of colour as another data stream. One approach to this could be to model the dominant object motion as well as the camera motion.

There is clearly much scope for further investigation into event detection in multi-modal sequences. One problem is being able to model the interactions between streams. The techniques used here model each stream independently so these interactions are not modelled. Clearly in most real situations this assumption of independence does not hold. A number of modifications to HMMs have been proposed to model these interactions [1] [4]. It is proposed to next carry out a comparision of different multi-modal sequence processing techniques on the same data sets. This will then provide a base line for the development of new techniques.

## REFERENCES


[1] S. Bengio, "An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition," in S. Becker, S. Thrun and K. Obermayer (eds.), **Advances in Neural Information Processing Systems, NIPS 15**, MIT Press, 2003.

[2] J. S. Boreczky and L. D. Wilcox, "A Hidden Markov Model framework for video segmentation using audio and image features," in **Proceedings of ICASSP**, 1998, vol. 6, pp. 3741–3744.

[3] P. Bouthemy, M. Gelgon and F. Ganansia, "A unified approach to shot change detection and camera motion characterization," **IEEE Trans. on Circ. and Systems for Video Technology**, vol. 9, no. 7, pp. 1030–1039, Oct 1999.

[4] M. Brand, N. Oliver and A. Pentland, "Coupled Hidden Markov Models for complex action recognition," in **Proceedings of IEEE CVPR97**, 1996.

| | Train set | Test set | Intragame | Intergame |
|---|---|---|---|---|
| Audio only | 70.6 | 64.1 | 64.2 | 64.0 |
| Motion only | 86.7 | 82.5 | 84.8 | 81.4 |
| Early fusion | 84.7 | 78.3 | 80.6 | 77.1 |
| Late fusion | 87.9 | 84.5 | 85.7 | 83.9 |

Table 3: A summary of percentage recognition rates for the training and test sets for all modes. The results for the test sets are averaged over the twelve non-diagonal values as shown in Figure 2 for each mode. The training results are an average of the diagonal values in Figure 2 for each mode. Intragame shows the average recognition rate when the training set and the test sets are from the same game. Intergame shows the average recognition rate when the training set and the test sets are from different games.


[5] P. Q. Dinh, C. Dorai and S. Venkatesh, "Video genre categorisation using audio wavelet coefficients," in **ACCV2002: The 5th Asian Conference on Computer Vision**, Melbourne, Australia, 2002.

[6] S. Dupont and J. Luettin, "Audio-Visual Speech Modelling for Continuous Speech Recognition," **IEEE Transactions on Multimedia**, vol. 2, no. 3, pp. 141 – 151, 2000.

[7] J. Huang, Z. Liu, Y. Wang, Y. Chen and E. Wong, "Integration of multimodal features for video scene classification based on HMM," in **IEEE 3rd Workshop on Multimedia Signal Processing**, 1999, pp. 53 – 58.

[8] J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, "On Combining Classifiers," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 20, no. 3, pp. 226–239, 1998.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in **Proc. Eurospeech '97**, Rhodes, Greece, 1997, pp. 1895–1898.

[10] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," **Journal of Visual Communication and Image Representation**, vol. 6, no. 4, pp. 348–365, Dec. 1995.

[11] S. Pfeiffer, S. Fischer and W. Effelsberg, "Automatic audio content analysis," in **Proc. 4th ACM Int. Conf. Multimedia**, Boston, Nov. 1996, pp. 21–30.

[12] L. Rabiner and B.-H. Juang, **Fundamentals of Speech Recognition**, PTR Prentice Hall, 1993.

[13] L. Xie, S.-F. Chang, A. Divakaran and H. Sun, "Structure analysis of soccer video with Hidden Markov Models," in **ICASSP**, 2002.

[14] G. Xu, Y.-F. Ma, H.-J. Zhang and S. Yang, "Motion based event recognition using HMM," in **Proceedings of ICPR**, Quebec, 2002.

[15] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in **Proc. ICME**, Tokyo, Japan, Aug 22-25 2001.

[16] Y. Wang, Z. Liu and J. Huang, "Multimedia content analysis using both audio and visual clues," **IEEE Sig. Processing Magazine**, vol. 17, no. 6, pp. 12–36, Nov 2000.