IDIAP RESEARCH REPORT

SPORTS EVENT RECOGNITION USING LAYERED HMMS

Mark Barnard * and Jean-Marc Odobez *

IDIAP–RR 05-07

JANUARY 2005

* Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) P.O. Box 592, CH-1920 Martigny, Switzerland. barnard@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-07

SPORTS EVENT RECOGNITION USING LAYERED HMMS

Mark Barnard and Jean-Marc Odobez

JANUARY 2005

Abstract. The recognition of events in video data is a subject of much current interest. In this paper, we address several issues related to this topic. The first one is overfitting when very large feature spaces are used and relatively small amounts of training data are available. The second is the use of a framework that can recognise events at different time scales, as standard Hidden Markov Model (HMM) do not model well long-term term temporal dependencies in the data. In this paper we propose a method combining Layered HMMs and an unsupervised low level clustering of the features to address these issues. Experiments conducted on the recognition task of different events in 8 rugby games demonstrates the potential of our approach with respect to standard HMM techniques coupled with a feature size reduction technique. While the current focus of this work focus is on events in sports videos, we believe the techniques shown here are general enough to be applied to other sources of data.

1 Introduction

With the recent growth in the amount of archive material there is a real need for systems capable of automatic content analysis and knowledge extraction. These applications would allow for structuring of video material in order to have efficient searching and retrieval of information. The problem of recognising particular events in video data pertains to many different areas, such as news and sports broadcasts, video surveillance and meeting annotation. Event recognition in video presents a number of significant problems. Firstly we have the problem of modelling temporal relations over a number of different time scales. For instance as well as modelling relations from one frame to the next we may also want to model the relations between longer term shots and events. There is also the problem of feature extraction and selection in video processing. This is still an open problem and currently the features used are selected or adapted to each specific application.

1.1 Temporal sequence modelling

One of the most common methods of modelling temporal sequences is *Hidden Markov Models* (HMMs) these are stochastic models with a discrete state space that can be trained using the *Expectation Maximisation* (EM) algorithm [1]. An HMM can be defined using two probability distributions the first $P(q_t = i|q_{t-1})$, where q_t is the state at time t, governs the transitions between states. The second $P(x_t|q_t)$, where x_t is the observation data at time t, is the probability of the data give the current state. HMMs have been successfully used in many different applications such as speech recognition, gene sequencing and gesture recognition. In general video processing tasks HMMs have been used with audio and video features in a scene classification task [2] and a video shot segmentation task [3]. Mccowan *et al* give a description of using various HMM topologies for recognition of events in meetings using multi-modal audio-video data [4]. In the specific area of sports video processing HMMs have been used to recognise events in basketball [5]. A good introduction to HMMs can be found in [1] and a through description of HMMs and their various extensions is available in [6]. While HMMs provide a good method of modelling temporal sequences they do suffer from overfitting when faced with a large number of parameters, long and complex temporal sequences and relatively small amounts of training data. HMMs also have difficulty modelling long term temporal relations in data. This is due to the state transition distribution which obeys the Markov assumption where the current state only depends on the the previous state.

2 Our approach

In this paper we propose a method using a Layered HMM (LHMM) to address the problems of modelling different time scales. In combination with this we propose to use unsupervised clustering of the data to address the problem of feature selection and dimension reduction in video data. In an effort to model long term relations in the data the Heirarchical HMM (HHMM) has been proposed [7]. These use HMMs at different levels in order to model data on different time scales. Xie *et al* use HHMMs to perform an unsupervised segmentation of *play* and *break* sequences in soccer videos[8] However as HHMMs use one large parameter space in order to do this they still suffer from the problem of overfitting and needing large amounts of training data. To segment the parameter space and increase the robustness to overfitting Layered HMMs (LHMM) were introduced [9]. This is an extention of HHMMs where each layer is trained independantly and the inferential results from the lower layer are used as data to train the layer above. We propose using a LHMM topology in combination with an unsupervised clustering of the features. First an unsupervised clustering of the data is done to produce a segmentation of the data into a number of mid-level clusters. The first layer of the LHMM, the Feature HMM (F-HMM) is used to produce a posterior probability for each of the mid-level clusters at each time t in the sequence. These probabilities are then used as features for the second layer of the LHMM, the higher level Event HMM (E-HMM), which then produces a probability of a higher level event at each time t. We would like to use the F-HMM to perform a dimension reduction of the feature space and so give more robust recognition in the higher level E-HMM. One problem we have is that there may be no obvious semantic decompostion of the higher level video events we are trying to recognise. This can be contrasted with decomposing group

actions in meetings into the individual actions of each person [10] or decomposing words into phonemes in speech recognition. In our case we use an unsupervised clustering of the data and then use this segmentation as a reduced mid-level set of features which can then be used for event recognition.

2.1 Unsupervised clustering

Here our goal is to segment the data X into different clusters. A cluster is represented by an HMM model M_i. M_i is a simple HMM with a single emitting state repeated several times to enforce a minimum duration constraint, one second in our experiments. The emission probability of that state is a GMM with parameters $\theta_i, N_i = |\theta_i|$; data belonging to the cluster i are denoted by D_i. θ_i and D_i are related, in the sense that θ_i are the parameters that fit the data D_i, while the D_i's can be computed from the data X and the parameters θ_i's using the standard HMM Viterbi decoding technique.

We would like to then find the optimal number of clusters k such that

$$k = \arg\max_k p(X, q_{best}|k), \tag{1}$$

where is q_{best} the path of the Viterbi decoding that gives the maximum likelihood. We use the following hierarchical clustering algorithm [11] to find the optimal solution. Starting with an over-segmentation of the data X, clusters are successively merged by replacing models M_a and M_b by the model M_{a+b} if the following criteria applies

$$\log p(D_{a+b}|\theta_{a+b}) \geq \log p(D_a|\theta_a) + \log p(D_b|\theta_b), \tag{2}$$

where $D_{a+b} = D_a \cup D_b$ and θ_{a+b} are the parameters fitting D_{a+b}. This criteria ensures an increase of the overall likelihood. An important point to note is that: $N_{a+b} = N_a + N_b$, which avoids the need to model the complexity of the models using BIC criteria for instance, with $N_{a+b} = |N_a|$ or $|N_b|$ for example. In this case it is well known that the right hand side of 2 would almost always be higher than the left hand side and no merging would occur. More details of this algorithm can be found in [11].

2.2 Connecting Layers in an LHMM

One of the principle problems in LHMMs is how to connect one layer of the model to the next, which means that the output of a layer can be used as an input feature to its higher layer. Here we will discuss the approach that has been taken to this problem. We define an observation sequence as: $X = x_1^T = \{x_1, x_2, \ldots, x_T\}$, where t is time and T is the length of the sequence. We can then think of this as two sequences, the past observation sequence, $x_1^t = \{x_1, x_2, \ldots, x_t\}$ and the future observation sequence, $x_{t+1}^T = \{x_{t+1}, x_{t+2}, x_{t+3}, \ldots, x_T\}$.

In the EM algorithm the forward variable α is defined as $\alpha(i,t) = P(x_1^t, q_t = i)$, this is the probability of having generated the past observation sequence and being in state i at time t. The backward variable β is defined as $\beta(i,t) = P(x_{t+1}^T|q_t = i)$, this is the probability that the future observation sequence will be generated given that we are in state i at time t. We also define the variable γ as $\gamma(i,t) = P(q_t = i|X)$, this is the probability of being in state i at time t given the entire observation sequence X [1].

In the original proposal for LHMMs by Oliver, Horitz and Garg [9] the layers are connect by using the values of α from the previous level as the observations for the next level. However recent work [10] has shown that a more principled and robust method of linking the layers of an LHMM is to use the values of γ. This approach has also recently been applied with success to speech recognition [12]. Here we will use the values of γ to link the two layers of the LHMM. This should provide a more accurate measure of the probability of the mid-level clusters as it uses all of the data X_1^T as opposed to α which is calculated using only the past observation sequence, X_1^t.

2.3 System Overview

The event recognition system we propose consists of an LHMM with two layers, a feature level HMM, F-HMM, and an event level HMM, E-HMM. In this system we use three sets of video features: motion, texture

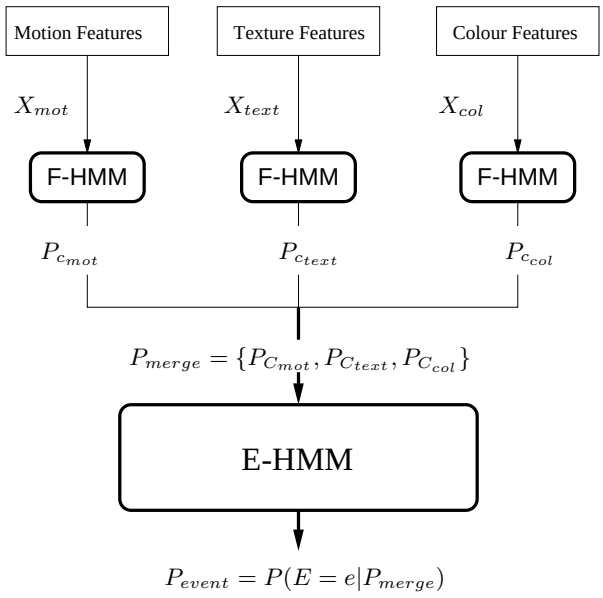

Figure 1: The proposed system with the F-HMM producing probabilities of the unsupervised clusters for each data stream and the E-HMM giving the probability of the annotated events.

and colour. So we produce a set of features for motion, texture and colour $x_t = \{x_{mot,t}, x_{text,t}, x_{col,t}\}$. An unsupervised clustering using the algorithm described above is then done for each feature set. There is a minimum duration of one second imposed on the clusters. This gives us a set of clusters for each feature set, with corresponding models, for motion M_{mot}, texture M_{text} and colour M_{col}. The F-HMM then produces a posterior probability $\gamma(t)$ for each of these models at each time t for each of the feature streams. This produces the following sets of probabilities: $P_{C_{mot}} = \{P(c^1_{mot} = c|X_{mot}), \ldots, P(c^{N_{mot}}_{mot} = c|X_{mot})\}$ for motion and similarly $P_{C_{text}}$ for texture and $P_{C_{col}}$ colour.

In the final stage the probability sets produced by each F-HMM are merged into a single high level feature set. This is then used as input to the E-HMM, which is trained using the supervised annotation of higher level semantic events. A graphical representation of this system can be seen in Figure 1.

3 Experiments

3.1 Events

We have selected three types of events we would like to recognise in rugby videos. The first are structural events these are common to most video material and describe the type of shot. Secondly we have play, non-play events these are common to most sports where the game is either being played or it is not being played. Lastly the events specific to the particular sport we are looking at, these events are dictated by the form and the rules of this sport. These events can be summarised as:

1. Structural events (6) - close up, person in a close up, long shot, miscellaneous, medium shot and medium shot low angle

2. Play events (3) - play, nonplay and replay

3. Action events (7) - running and passing, maul, line-out, kick, penalty, scrum and try

In the following experiments we have made no assumptions about any heirarchy in these sets of events. Currently we treat these as three separate and independant annotations of the same data, though in future work we will consider the interactions between them.

3.2 Features

The motion features were used in this experiment to characterise the dominant motion model over the entire image field of view [13]. In calculating the texture features the image is divided into 20 equal rectangles and then an edge direction histogram for each region is calculated. The colour feature is based on a playfield segmentation algorithm developed in previous work [14] and calculates the percentage of playfield in each region of the image.

3.3 Data sets and evaluation protocol

The data used in these experiments consists of 7 half games. Each half game is a video file approximately 45 to 50 minutes long. We divided this data into two sets, one for training and validation, five games, and the other for testing, two games. This data was then annotated by hand with the high level structural, play and action events. In the results we present we have used the frame recognition rate as a measure of performance. This is given by dividing the number of frames correctly classified by the total number of frames tested. This was chosen over other common measures such as Word Error Rate because in this application we want to be sure we correctly recognise the timing of events as well, this allows for searching and retrieval of the data.

We tested the performance of our method against using the raw features and also against a common method of dimension reduction *Principle Component Analysis* (PCA). Using PCA we reduced the size of the original feature vector from 92 to 37 with these 37 features still accounting for 90% of the variance in the original data. Using the unsupervised clustering we reduced the final feature vector size in the proposed LHMM system to 35. We trained all three models, HMMs, HMMs + PCA and LHMMs with clustering, on the annotated data and then adjusted the word insertion penalty and the minimum duration using the training set. All models were trained with a single state and 20 gaussian mixtures

3.4 Results and discussion

Method	Frame Rec Rate	
	Training set	Test set
Standard HMMs	0.83	0.64
Standard HMMs + PCA	0.82	0.57
Layered HMM + clustering	0.76	0.67

Table 1: Results for structural events for standard HMMs, HMMs trained on data after PCA and Layered HMMs using unsupervised clustering

Method	Frame Rec Rate	
	Training set	Test set
Standard HMMs	0.78	0.70
Standard HMMs + PCA	0.76	0.67
Layered HMM + clustering	0.79	0.79

Table 2: Results for play events for standard HMMs, HMMs trained on data after PCA and Layered HMMs using unsupervised clustering

It can be seen from the results shown in Table 1, 2 and 3 that the proposed technique offers clear improvements for all three classes of events. The robustness of the our method can be seen by comparing the performance on the training and the testing set. It is clear in many cases that standard HMM approach is prone to overfitting in this task. We can also see that our method is clearly a more robust form of feature space reduction than the standard PCA approach. Indeed even with the reduction in the feature space size the traditional

Method	Frame Rec Rate	
	Training set	Test set
Standard HMMs	0.70	0.69
Standard HMMs + PCA	0.69	0.57
Layered HMM + clustering	0.73	0.74

Table 3: Results for action events for standard HMMs, HMMs trained on data after PCA and Layered HMMs using unsupervised clustering

HMM models still show signs of overfi tting. While these result are very encouraging we believe their is potential for exploiting the ability of LHMM to model events on different time scales in order to further improve the results.

References

[1] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

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

[3] John S. Boreczky and Lynn 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.

[4] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear)*, 2004, To appear.

[5] Gu Xu, Yu-Fei Ma, Hong-Jiang Zhang, and Shiqiang Yang, "Motion based event recognition using HMM," in *Proceedings of ICPR*, Quebec, 2002.

[6] Kevin Murphy, *Dynamic Bayesian Networks: Representation, inference and learning*, Ph.D. thesis, UC Berkeley, 2002.

[7] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden markov model," *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.

[8] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models," in *Proc. ICME*, July 2003.

[9] N. Oliver, E. Horitz, and A. Garg, "Layered representations for leanring and inferring offi ce activity from multiple sensory channels," in *Proc. ICMI*, October 2002.

[10] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling Individual and Group Actions in Meetings With Layered HMMs," IDIAP-RR 33, IDIAP, Martigny, Switzerland, 2004, submitted for publication.

[11] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[12] H. Bourlard, S. Bengio, M. Magimai Doss, Q. Zhu, B. Mesot, and N. Morgan, "Towards using hierarchical posteriors for flexible automatic speech recognition systems," IDIAP-RR 58, IDIAP, 2004.

[13] 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.

[14] M. Barnard and J.M. Odobez, "Robust playfi eld segmentation using map adaptation," in *Proc. 17th ICPR (ICPR 2004)*, Cambridge, United Kingdom, 2004.