THIS PUBLICATION APPEARED IN THE IEEE ICME 2005 CONFERENCE

SPORTS EVENT RECOGNITION USING LAYERED HMMS

Mark Barnard and Jean-Marc Odobez

IDIAP Research Institute
Rue du Simplon 4
Martigny Switzerland

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 7 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 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 systems 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. Feature extraction and selection is a second problem in video processing. Often, the recognition of a particular event in this domain is better addressed by designing a highly specialised feature extractor. In this paper, we focus our study on a more generalised approach to event detection in video data.

The authors acknowledge financial support provided by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. This work was also supported by the EU-IST project LAVA

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 by probability distributions; the first $P(q_{t-1} = j|q_{t-1})$, where q_{t-1} is the state at time $t-1$, governs the transitions between states. The second $P(x_t|q_t = i)$, where x_t is the observation data at time t, is the probability of the data given 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* describe using various HMM topologies for recognition of events in meetings using 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 thorough 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 an effort to model long term relations in the data Hierarchical HMMs (HHMMs) have 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 the parameter space of HHMMs is still large, they suffer from the problem of overfit-

ting and needing large amounts of training data. To reduce the size of the parameter space and increase the robustness to overfitting Layered HMMs were introduced [9]. Layered HMMs can be seen as a variant 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. While less powerful at modeling long term temporal relationships than HHMMs, Layered HMMs offer a way of reducing the dependency of training with respect to the input feature space.

In this paper we propose a method using a Layered HMM 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. The first layer of the Layered HMM, 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. This layer is built by using an unsupervised clustering and segmentation of the training data, this is described in section 2.1.

These probabilities are then used as features for the second layer of the Layered HMM. This second layer is trained using the output of the first layer. This is supervised training using the higher level events we want to recognise. So the higher level Event HMM (E-HMM) produces a probability of a higher level event at each time t. An overview of this system can be seen in Figure 1.

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 training data $\mathcal{D}$ 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. The emission probability of that state is a Gaussian Mixture Model with N_i mixtures and parameters θ_i. The segments of data belonging to the cluster i are denoted by D_i. As with standard clustering θ_i and D_i are related more specifically, θ_i are the parameters that fit the temporal data D_i, while the D_i's can be computed from the data $\mathcal{D}$ and the parameters θ_i's using the standard HMM Viterbi decoding technique.

We use the following hierarchical clustering algorithm [11] to find the optimal solution. The overall goal in the clustering segmentation is to find the optimal number of clusters k such that

$$\hat{k} = \arg\max_k p(\mathcal{D}, q_{best}|k), \quad (1)$$

where is q_{best} the path of the Viterbi decoding for which the maximum data likelihood. 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

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

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 in this algorithm [11] the complexity of the merged model M_{a+b} is kept similar to that of the sum of the individual ones M_a and M_b by letting $N_{a+b} = N_a + N_b$. This avoids the need to model the complexity of the models using BIC criteria for instance.

2.2. Connecting Layers in an Layered HMM

One of the issues in Layered HMM modeling is how to connect one layer of the model to the next, that is what 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.

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 Layered HMMs by Oliver, Horitz and Garg [9] the layers are connected by using the value $P(q_t = i|x_t)$ 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 Layered HMM is to use the value $P(q_t = i|x_1^t)$. Performance was further improved by the use of the posterior probability 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 Layered HMM. 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

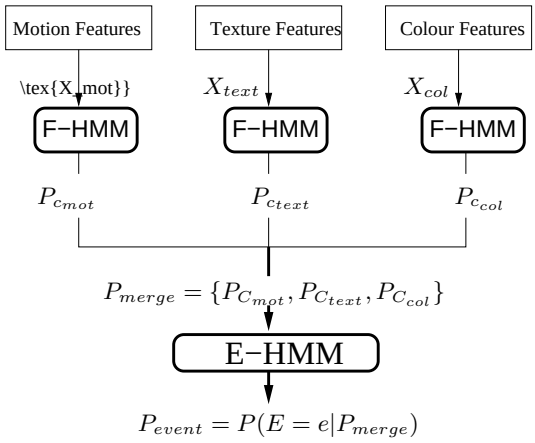

Fig. 1. The proposed system with the F-HMM producing probabilities of the unsupervised clusters for each data stream and the E-HMM giving the probability of events.

the past observation sequence, X_1^t. This method does, however, require batch processing as the entire sequence must be processed by the F-HMM before data is available to the E-HMM.

2.3. System Overview

The event recognition system we propose consists of an Layered HMM 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 and colour, $x_t = \{x_{mot,t}, x_{text,t}, x_{col,t}\}$. An unsupervised clustering using the algorithm described above is then applied separately on each feature set. We enforce a minimum duration of one second on the cluster segments. 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 layer 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 streams of probabilities: $P_{C_{mot}} = \{P(c_{mot}^1 = c|X_{mot}), \ldots, P(c_{mot}^{N_{mot}} = c|X_{mot})\}$ for motion and similarly $P_{C_{text}}$ for texture and $P_{C_{col}}$ colour.

In the second 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.

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 which are common to many sports video material and describe the type of shot: medium shot and medium shot low angle close up, person in a close up, long shot, miscellaneous. Secondly, we have play events: play, nonplay and replay. Lastly action events that are specific to the particular sport we are looking at. These events are dictated by the form and the rules of the selected sport. In our work, for rugby we defined: 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 used in our experiments characterise the dominant motion model over the entire image field of view [13]. In the texture case, the image is divided into 20 equal rectangles and then an edge direction histogram for each region is calculated. The colour feature, are based on a playfield segmentation algorithm developed in previous work [14] and consists of the percentage of playfield in each of the 20 regions of the image. The size of the feature vectors for motion, texture and colour are 9, 63 and 20 respectively.

3.3. Data sets and evaluation protocol

The data used in these experiments consists of 7 half games of approximately 45 to 50 minutes. 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.

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 four models, HMM, HMM-PCA, HMM-PCA-R and LHMM 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.

In the results we present we have used the frame recognition rate as a measure of performance for the play and structural events. This is given by dividing the number of frames correctly classified by the total number of frames tested. In the case of action events, however, as the events are very unbalanced we want to report performance based on event recognition rather than frame recognition. We thus introduce a new measure based on the edit distance between

the groundtruth sequence and the recognised sequence with an added constraint that in order to match the events must co-occur in time. The event based precision and recall are then calculated based on the alignment by the edit distance optimisation. Precision is given by E_{corr}/E_{rec} and recall by E_{corr}/E_{ground}, where E_{corr} is the number of correctly recognised events, E_{rec} is the total number of recognised events and E_{ground} is the total number of events in the groundtruth.

3.4. Results and discussion

Method	Training set	Test set
HMM	0.83	0.40
HMM-PCA	0.80	0.59
HMM-PCA-R	0.82	0.57
Layered HMM	0.76	0.67

Table 1. Frame recognition rate for structural events.

Method	Training set	Test set
HMM	0.78	0.70
HMM-PCA	0.77	0.70
HMM-PCA-R	0.76	0.67
Layered HMM	0.79	0.79

Table 2. Frame recognition rate for play events.

Method	Training set	Test set
HMM	0.70	0.69
HMM-PCA	0.68	0.55
HMM-PCA-R	0.69	0.57
Layered HMM	0.73	0.74

Table 3. Frame recognition rate for action events.

Method	Micro average		Macro average	
	Rec	Prec	Rec	Prec
HMM	0.40	0.47	0.18	0.28
HMM-PCA	0.37	0.56	0.36	0.49
HMM-PCA-R	0.26	0.58	0.20	0.45
Layered HMM	0.52	0.69	0.41	0.49

Table 4. Micro and macro precision and recall rates for action events.

It can be seen from the results 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 frame recognition rates on the training and the testing set in Tables 1, 2 and 3. It is clear in many cases that standard HMM approach is prone to overfitting in this task, and 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 HMM models still show signs of overfitting.

Tables 4 and 5 provide the macro average (the average of recall and precision computer per class) and the micro average (the average weighted by class size). They confirm that the Layered HMM is performing better. However the rates appear to be quite poor with approximately half of the events recognised. Better results may have been obtained by specifically tailoring features for this application. Finally these result are very encouraging we believe their is potential for exploiting the ability of Layered HMM to model events on different time scales in order to further improve the results .

4. 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 classification based on HMM," in *IEEE 3rd Workshop on Multimedia Signal Processing*, 1999.

[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.

[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* , vol 27, No. 3 March 2005.

[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 learning and inferring office 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,

[11] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE ASRU 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 playfield segmentation using map adaptation," in *Proc. 17th ICPR (ICPR 2004)*, UK.