HMM Model Selection Issues for Soccer Video

Mark Baillie, Joemon M. Jose, and Cornelis J. van Rijsbergen

Department of Computing Science, University of Glasgow,
17 Lilybank Gardens, Glasgow, G12 8QQ, UK
{bailliem, jj, keith}@dcs.gla.ac.uk

Abstract. There has been a concerted effort from the Video Retrieval community to develop tools that automate the annotation process of Sports video. In this paper, we provide an in-depth investigation into three Hidden Markov Model (HMM) selection approaches. Where HMM, a popular indexing framework, is often applied in a ad hoc manner. We investigate what effect, if any, poor HMM selection can have on future indexing performance when classifying specific audio content. Audio is a rich source of information that can provide an effective alternative to high dimensional visual or motion based features. As a case study, we also illustrate how a superior HMM framework optimised using a Bayesian HMM selection strategy, can both segment and then classify Soccer video, yielding promising results.

1 Introduction

Live televised sporting events are now common place, especially with the arrival of dedicated digital channels. As a result, the volume of Sports video produced and broadcasted has increased considerably over recent years. Where such data is required to be archived for reuse, automatised indexing [2,3,5,8,9] is a viable alternative to the manual labour intensive procedures currently in practise. To date feasible solutions have not been developed. Current advancements, mainly the automatic identification of low level semantic structures, such as shot boundaries [3], semantic units [5,9] and genre classification [8] can reduce both the time and workload for manual annotation. Also, recognition of such low level structure is the basis for which further processing and indexing techniques can be developed. For example, labelling of low level segments can enable domain specific indexing tools such as exciting event detection [2] to be enhanced, utilising prior knowledge of content.

The difficulty with indexing Soccer video is that unrelated semantic components can contain visually very similar information, resulting in accuracy problems. For example, it is not uncommon for advertisements to display Sport sequences during televised events, to boost marketing appeal of a product, a potential source for error. However, audio is a rich, low dimension alternative to visual information that can provide an effective solution to this problem.

In this paper we model audio content using the Hidden Markov Model (HMM), a popular indexing framework. The main thrust of this research is

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 70–78, 2004.

to provide an in-depth investigation into HMM model selection, where HMM is largely applied in an ad hoc manner for video content indexing [5,8,9]. We also investigate what effect poor selection can have on future indexing accuracy.

The remainder of this paper is structured as follows. In Section 2, we identify the potential factors that influence the application of a HMM. We then formally investigate three model selection strategies, in Section 3. As a case study, in Section 4, we illustrate how an extended HMM framework for segmentation and classification of Soccer video, can be optimised using model selection, yielding promising results. Finally, we conclude our work in Section 5.

2 Hidden Markov Model Issues

HMM is an effective tool for modelling time varying processes, belonging to a family of probabilistic graphical models able to capture the dynamic properties of temporal data [7]. Similar static representations, such as the Gaussian Mixture Model (GMM), do not model the temporal properties of audio data, hence the popularity of HMM in the fields of Speech Recognition [4,7], temporal data clustering [6,7] and more recently Video Retrieval [2,3,5,8,9]. An important issue when employing a continuous density HMM framework is model selection [6,4, 7]. For example, a crucial decision is the selection of both an appropriate number of hidden states and (Gaussian) mixture density estimation per state. Accurate segmentation and classification is dependent on optimal selection of both these parameters. An insufficient number of hidden states will not capture enough detail, such as data structure, variability and common noise, thus loosing vital information required for discrimination between groups. A greater number of hidden states would encapsulate more content, though precise and consistent parameter estimation is often limited by the size and quality of the training data. As the number of parameters increase, so does the number of training samples required for accurate estimation. Larger more enriched models require a greater volume of training data for precise parameter estimation. A further problem with complex models is overfitting. HMMs, specifically designed to discriminate between content, can become too detailed and begin to mirror nuances found in unrelated groups, deteriorating classification accuracy.

HMM application for Video Retrieval has so far been ad hoc, with little investigation into model selection and the potential side effects on system performance. In the literature, a common theme is to apply domain knowledge or intuition for HMM model selection. Such application includes shot boundary detection [3], news video segmentation and classification [5], TV genre labelling [8] and 'Play' or 'Break' segmentation [9] for Soccer video. This strategy can be helpful when matching a known number of potential states found in the data, such as shot segmentation [3]. However, there has been little research into how suitable this strategy is when applied to broad content classes found in video. For example, Wang et. al. [8] employ the same number of hidden Markov states for modelling entire video genre such as Sport and News, ignoring differences in the underlying structure found in each separate domain.

Eickeler et. al. [5], apply domain knowledge to News Broadcasts, building a superior HMM based on a preconceived topology. Each state of a superior HMM is represented by a simple HMM that models a broad content class found in News video. However, there is no investigation into model selection for these simple HMMs. Xie et al [9] segment and classify 'Play' and 'Break' segments for Soccer video, by using HMMs to model motion and colour distribution statistics. 'Play' segments correspond to camera shots that track the flow of the game. To model both segments, the authors use a series of simple HMM models, with a varying number of hidden states. For segmentation and classification, the output from each model is then combined using a dynamic programming (DP) algorithm, itself a first order Markov process. In fact, this application ignores the temporal properties of the HMM, suggesting a simpler classifier such as the Gaussian Mixture Model, applied in conjunction with the DP algorithm, may be as effective.

3 HMM Model Selection

The main goal of model selection is to choose the simplest possible model without a deterioration in performance. This is especially important given the difficulty and practicality of generating large, varied training sets. In this Section, we investigate three model selection strategies and what effect each has on classification performance. The three selection strategies are: an exhaustive search approach, the Bayesian Information Criterion (BIC) [4,6] and the Akaike Information Criterion (AIC) [1] (formulae can be found in references). Exhaustive search, a simple linear search algorithm, involves training and testing a series of HMMs, where the parameter in question is iteratively increased until a stopping threshold is reached. For each iteration, the predictive likelihood of a HMM generating a test sample is calculated, also known as the out of sample log-likelihood. Using a stopping criteria on the predictive likelihood score is important. For example, increasing the number of states will also increase the predictive likelihood until each training sample is eventually modeled by its own unique hidden state.

The two remaining strategies are BIC and AIC, both popular in the Statistical literature. Each strategy penalises the predictive likelihood with a term that is derived from the number of parameters in the model. The major difference between approaches, is the derivation of this penalty term. The penalty term for AIC, only accounts for the number of free parameters in the HMM, while the BIC penalty term also factors in the amount of training data available. Smaller training samples will generate larger penalty scores, hence the advantage in predictive likelihood found with more complex models is eventually outweighed by this penalty term. We then assume the optimal model is found at the maximum predictive likelihood score, avoiding the need to threshold.

3.1 Data Set

To evaluate each strategy, we generated a data set of 12 games ranging between 2 to 3 hours in length. We manually labelled the audio into three main semantic content classes found in Soccer video; 'Game', 'Studio' and 'Advertisement'.

'Studio' segments contain an introduction plus pre and post match discussion and analysis of the live game, usually set inside a controlled soundproof studio. 'Game' segments consist of the live match, where the soundtrack contains a mixture of both commentary and vocal crowd reaction, alongside other environmental sound such as whistles, drums and clapping. 'Advert' segments can be identified by the almost chaotic mixture of highly produced music, voice and sound effects. Segmentation and labelling of these low level segments is beneficial, especially for reducing indexing errors during higher level tasks. For example, identifying the boundaries of a 'Game' segment is vital before event detection [2]. A decrease in precision would occur if the data was not pre-segmented and labelled. Similar information from unrelated content such as music or sound effects, can then be wrongly identified as a key event.

3.2 Number of Markov States Selection

A series of HMMs were implemented, modelling the 'Game', 'Studio' and 'Advert' content classes. The audio stream for each file was parameterised using 14 Mel-Frequency Cepstral coefficients (MFCC) with an additional Log Energy measurement [7]. MFCC coefficients are specifically designed and proven to characterise speech well. MFCC has also shown to be both robust to noise and useful in discriminating between speech and other sound classes [2,4].

For each class, a series of ergodic, continuous density HMMs [7] with increasing number of states ranging from 1 to 20, were iteratively implemented. Each model was first generated from a training sample, then the predictive likelihood score was calculated on a separate test set. Both labeled data samples were generated from the same pool of data and after one complete run, each sample was randomly changed. This was repeated 15 times to achieve a true reflection of the model generation process, limiting the effect of unusually good or bad runs.

Importantly, each HMM was assigned a singular Gaussian density per hidden state. An informal investigation using synthetic data, indicated that it was more important to identify the correct number of states first, to avoid searching through an unnecessary large number of hidden state and mixture combinations. For example, 100 HMMs of different state and mixture combination were implemented using data generated by a 6 state HMM, with 6 mixtures per state, Figure 1(a). Using the exhaustive search approach, increasing the number of mixtures did not effect correct hidden state number identification. As a result, we could first identify the optimal number of hidden Markov states for each content class, implementing a singular density function per state. Then in a separate step, the optimal number of Gaussian density components could be found, reducing the number of parameter combinations to be implemented.

Using the 'Game' class as an example, Figure 1(b) displays the mean of the 15 initialisations, for all selection strategies. For the exhaustive search approach, the predictive likelihood increases as a new hidden Markov state is added to the model. There is a rapid rise that levels off between 14 to 20 states, suggesting the model was beginning to overfit the training data. A stopping threshold, empirically determined using synthetically generated data, Figure 1(a), was reached

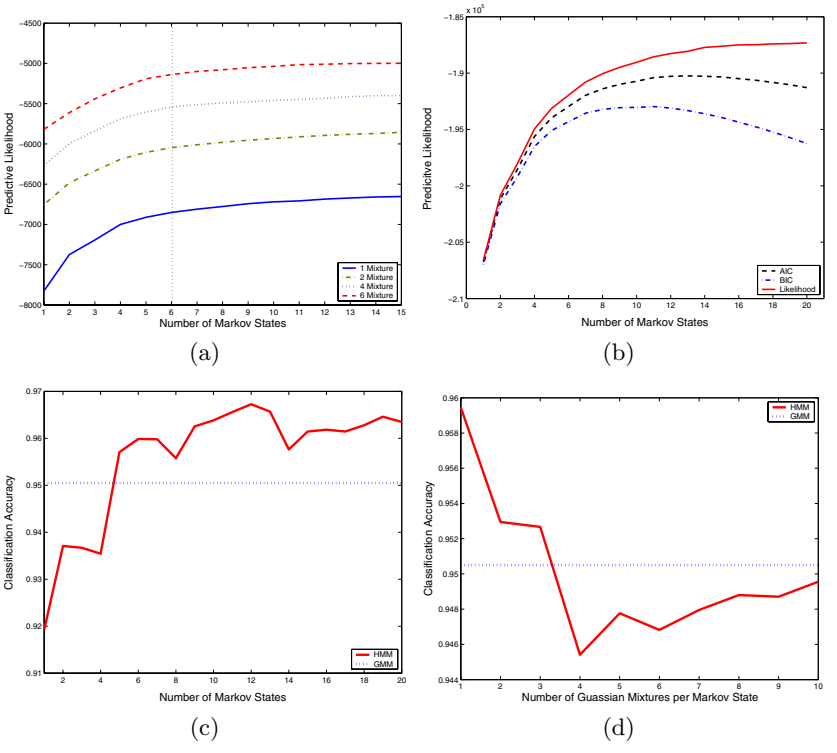

Fig. 1. (a) The predictive likelihood scores for HMMs with increasing state and mixture component number. (b) A comparison of model selection strategies for hidden state selection. Notice, both the AIC and BIC scores peak, while the predictive likelihood score continues to increase. (c) Classification accuracy versus number of hidden Markov states. (d) Classification accuracy versus the number of Gaussian mixture components.

when adding a 14^{th} state. For the BIC strategy, the predictive likelihood also increased dramatically but peaked and then tailed off. The maximum BIC score was found to be 9 states. For the AIC strategy, a similar pattern occurred, where the maximum AIC score was found at 12 states. There was a similar trend for the remaining two content groups. BIC selected the simplest model followed by AIC, then the exhaustive search method.

We also evaluated what effect iteratively adding hidden Markov states had on classification accuracy, Figure 1(c). As a comparison, the simpler GMM classifier [7], which does not model time as a property, was used as a baseline. The mean classification accuracy gradually increased as new hidden states were added to the HMM. After the 5^{th} hidden state was added, the HMM began to outperform the GMM classifier. A 12 state HMM was found to be optimal for this content class, the same model selected using the AIC strategy. A similar pattern emerged for the remaining content classes. An improvement in classifi-

cation accuracy over the baseline GMM was recorded, when a certain number of states were added to the HMM.

3.3 Number of Gaussian Mixtures per Markov State

The same implementation issues arise with the selection of mixture components that model the emission densities per hidden Markov state. For example, speech recognition systems have identified that HMMs with multiple Gaussian mixtures perform better than those with a singular density function [4,7]. A mixture of Gaussian components can model the multi-modal emission densities that represent variation found in speech. However, selecting too many mixture components can result in overfitting. Thus, we repeated the previous experiment, this time implementing HMMs with increasing Gaussian mixture components per state.

For each strategy, each content class was modeled with mixtures ranging from 1 up to 10, fixing each HMM with the optimal number of hidden states identified in the previous section. For example, for one content class, 3 HMMs were implemented using the optimal number of states identified by each selection strategy. To limit overfitting further. The covariance matrices were constrained to be diagonal for each individual mixture, reducing the number of free parameters. Each model setting was initialised 15 times, changing the data samples randomly after a complete run. Our findings again indicated that the BIC strategy selected the simplest model followed by AIC. The exhaustive search strategy again selected the more complex HMMs.

We also analysed what effect iteratively adding Gaussian mixtures per model had on classification accuracy, Figure 1(d). From our results, we discovered a decrease in classification accuracy as mixtures were added to a singular density HMM. This trend was consistent across all strategies and for all content classes. Figure 1(d), is an illustration of a 9 state HMM for the 'Game' class, as the number of mixture components is iteratively increased. Classification accuracy decreases until 4 states are added, with a small reverse in trend afterwards. After three mixtures, the model performance became poorer than that of the GMM. This result was mirrored across the remaining two content classes and could be indicative of both poor parameter estimation given increased model complexity, as well as overfitting. To summarise. A singular density HMM produced the best classification accuracy when compared to the same HMM with multiple mixture components.

3.4 Optimal HMM Model Evaluation Experiment

In the previous section, we identified 3 optimal HMMs for each content class, using three selection strategies. Next, these HMMs were formally compared over a new test set, using a baseline GMM classifier for comparison. The test set was approximately 2 hours in length, divided into 10 second observation sequences, labelled into each content class. For all content classes, a HMM was first generated from the labeled data used in the previous section. The HMM was then tested on the new sample. For each strategy, each new individual sequence was

Table 1. Confusion matrix. The % of correctly classified observations are in bold.

	Classification (%)												
Correct	Game				Studio				Advert				
Class	LIK	BIC	AIC	GMM	LIK	BIC	AIC	GMM	LIK	BIC	AIC	GMM	Total
Game	**89.6**	**92.7**	**90.4**	**90.4**	1.8	1.1	1.0	2.9	8.5	6.2	8.6	6.7	*100%*
Stud	4.6	5.2	5.1	2.9	**89.1**	**86.8**	**87.6**	**90.3**	6.2	8.0	7.2	6.8	*100%*
Advt	1.1	1.0	1.1	1.4	3.5	3.4	3.0	3.7	**95.4**	**95.6**	**95.9**	**94.9**	*100%*

assigned to the content class that produced the highest HMM likelihood score, found using the Viterbi decoding algorithm [7].

The results in Table 1, indicated no significant difference in terms of classification accuracy across all selection strategies, and across each content class. Overall, the 'Studio' classifier indicated the worst performance, where the majority of false classifications were samples with speech containing background noise, wrongly labelled as 'Game' or 'Advert'. False classification from the 'Game' class again included sequences containing speech. These observations contained little or no environmental sound associated with the 'Game' class, resulting in misclassification. Samples containing music played inside the stadium, or other peculiarities such as tannoy announcements, were also wrongly labelled into the 'Advert' class. These sound events reflected similar content found in the 'Advert' class. The 'Advert' HMM produced the highest classification accuracy for all selection methods, where the majority of false classifications were labeled into the 'Studio' category. These errors were typically clean speech samples.

Given that the BIC selection criterion chose the simplest HMMs overall, there was no obvious detriment in performance. In fact the HMM selected by BIC for the 'Game' class, produced the highest classification accuracy. However, the same selection strategy resulted in the lowest classification accuracy for the 'Studio' group. Interestingly, for the same content class the baseline GMM classifier recorded the best result. In fact, across all content classes, the GMM displayed comparable results when compared to the HMM.

3.5 Discussion

From experimentation, we illustrated the importance of model selection, where a gain in performance can be found when selecting HMMs methodically. For many Video indexing applications of HMM, this type of approach is not adopted [5, 8,9], highlighting optimisation issues for each system. Selecting too few or too many hidden states can produce poor classification performance, as shown from the experimentation of three model selection techniques.

The BIC method selected the simplest HMMs without significantly decreasing classification accuracy. In some cases, displaying a higher classification accuracy than more complex HMMs. Also, the BIC strategy has an obvious advantage over an exhaustive search approach. The BIC penalty term creates a maximum peak in the predictive likelihood score. We assume this maxima to be the optimal

HMM. Hence, to find an optimal solution. The number of HMMs required to be implemented can be reduced by avoiding an iterative addition of parameters. For example, a bisection search strategy such as a Newton-Raphson could be implemented to find the maximum BIC score.

From experimentation, an important discovery was the effect increasing the number of mixture components had on classification accuracy. Adding further Gaussian mixtures to a singular density HMM, created a detrimental effect. Increasing the complexity resulted in poor parameter estimation and overfitting. In most cases, after two or more mixtures were added, the baseline GMM recorded better results. In fact, for the task of classification, the HMM framework did not perform significantly better than the GMM overall. For this problem, GMM has been shown to be as effective when compared to the more complex HMM.

4 A Segmentation and Classification System

In this section, our aim is to segment and then classify Soccer video files using audio information alone. We present a case study, illustrating how optimally selected HMMs using BIC, can be integrated into a superior HMM framework [5]. This combination scheme utilises both domain knowledge as well as statistical model selection, where each optimised HMM becomes a single state in a unified HMM topology. This superior HMM allows for an entire video file to be segmented, classifying semantic segment units in a single pass. The advantage of applying this decision process is the ability to incorporate the temporal flow of the Video into the segmentation process, limiting error. For example, restricting movement from the 'Advert' to 'Game' segments can be mirrored in the state transition matrix in the superior HMM. Also, an input and output state, to note the beginning and end of each video file are included.

To evaluate this technique, given the limited data set, we applied a 'leave one out cross validation'. 11 complete video files were used for model training. The 'held' out video was then used to evaluate the superior HMM. This procedure was repeated, holding out each video in turn, until segmentation and classification was achieved for all videos in the data set. We indexed all 12 video files using the Viterbi decoding algorithm, where each one second is assigned to a state in the superior HMM that represented a specific content class. An ambiguity window of 2 seconds was allowed for each segment change when comparing the indexed files with the manually generated truth data. This was to limit small alignment errors between the ground truth and the model output.

The majority of segment boundaries were identified with 95.7% recall and 89.2% precision. 97.9% of the segments were correctly labeled. Even allowing for the ambiguity window. Those segment changes that were not picked up correctly were largely due to alignment errors, where the detected boundary was missed by a few seconds. False detections for segment change mostly involved wrongly identified segment transition between 'Studio' to 'Game' segments or vice versa. For example, false boundary changes were marked during a 'Game' segment where there was a decrease in crowd sound. A simple solution to this

problem would be to add a state duration element into the HMM framework. One complete 'Game' segment spans approximately 45 minutes. Incorporating a time distribution could avoid false classifications, especially during quiet spells in a 'Game' segment.

5 Conclusions and Future Work

In this paper, we investigated three HMM model selection strategies, examining factors that can effect the application of a HMM framework. We found the BIC selection strategy to be the most effective. By then incorporating optimal HMMs into a unified framework, we then illustrated how a superior HMM can be applied to both segment and classify the low level structure of Soccer video, yielding promising results. Labeling was achieved by modelling underlying audio patterns found in each semantic unit.

Intended future work will include the extension of the superior HMM framework to include visual, motion and textual information sources. Another active area of interest will be incorporating the classification of smaller sub-groups such as crowd cheering for event detection [2], music and speech. Thus extending the HMM framework to include a more complete topology for the annotation of live Soccer broadcasts. Finally, we wish to compare this system against other frameworks, a requirement highlighted during experimentation.

Acknowledgements. The authors would like to thank Prof. Mark Girolami and Vassilis Plachouras.

References

1. H. Akaike. A new look at the statistical model identification. In *Trans. Automatic Control*, volume AC-19, pages 716–723. IEEE, Dec 1974.
2. M. Baillie and J. M. Jose. Audio-based event detection for sports video. In *CIVR2003*, pages 300–310, Il, USA, July, 2003.
3. J. S. Boreczky and L. D. Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *ICASSP*, pages 3741–3744, Seattle, May 1998. IEEE.
4. S. S. Chen and R. A. Gopinath. Model selection in acoustic modeling. In *Proceedings of Eurospeech-99*, Hungary, September 1999. Eurospeech.
5. S. Eickeler and S. Muller. Content-based video indexing of tv broadcast news using hidden markov models. In *ICASSP*, Phoneix, USA, 1999. IEEE.
6. C. Li and G. Biswas. A bayesian approach to temporal data clustering using hidden markov models. In *ICML*, pages 543–550, California, 2000.
7. L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.
8. Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. In *IEEE Signal Processing Magazine*. IEEE, 2000.
9. L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *ICASSP*, 2002.