Temporal Post-processing of Decision Tree Outputs for Sports Video Categorisation

Edward Jaser, William Christmas, and Josef Kittler

Centre for Vision, Speech and Signal Processing, University of Surrey
Guildford GU2 7XH,UK
Tel.: +44 (0)1483 689294, Fax: +44 (0)1483 686031
{E.Jaser,J.Kittler,W.Christmas}@eim.surrey.ac.uk

Abstract. In this paper, we describe a multistage decision making system to deal with the problem of automatic sports video classification. The system is founded on the concept of cues, i.e. pieces of visual evidence, characteristic of certain categories of sports that are extracted from key frames. The main decision making mechanism is a decision tree which generates hypotheses concerning the semantics of the sports video content. The final stage of the decision making process is a Hidden Markov Model system which bridges the gap between the semantic content categorisation defined by the user and the actual visual content categories. The latter is often ambiguous, as the same visual content may be attributed to different sport categories, depending on the context. We tested the system using two setups of HMMs. In the first, we construct and train an HMM model for each sport. A post-processing step is needed in this setup to combine the outcomes of the individual HMMs. In the second setup, we eliminate the need for post-processing by constructing a single HMM with each node representing one of the sports we want to detect. Comparing the results obtained from both setups showed that a single HMM delivered the better performance.

1 Introduction

In this paper we consider the problem of automatic sports video categorisation. This problem arises during multidisciplinary events such as Olympic Games where huge volumes of video material are recorded, with the content randomly switching from one discipline to another. A coarse automatic annotation in terms of sport identity would aid the production of event summaries for news cast and other applications.

Much research in the field of multimedia analysis and retrieval is targeting the domain of sport videos. The reason is that most sport videos have a well-defined content structure and official rules and procedures compared to videos from other domains. Moreover, most sporting events take place in one location. That means only a limited number of cameras are needed to cover the play area and capture the event. Therefore, a set of characteristic views recorded by those cameras can be defined and associated with the events. Figure 1 gives an

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 495–503, 2004.

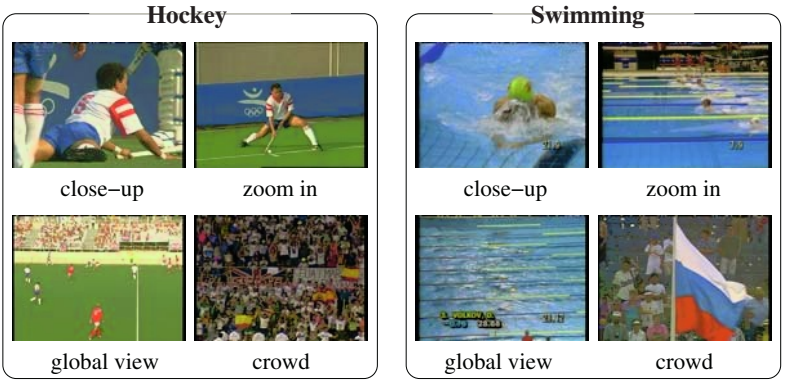


Fig. 1. Sport views

example of wide variety of views that exist in two sport disciplines, swimming and hockey.

Other work, specific to some form of sports annotation, include [3] in which the authors addressed the problem of parsing the content of football video programs. They used domain knowledge about football to construct an a priori model to aid the task of classifying key-frames of each shot to a predefined set of events. Xu et al [12] also addressed the problem of segmenting football videos into two basic semantic units "play" and "break". Addressing football videos analysis and summarisation as well, Ekin et al [2] proposed a fully automated system using both cinematic and object-based features. Chang et al [1] proposed a statistical method for the automatic extraction of predefined highlight segments in a baseball game video using an HMM built for each class of highlight. HMMs were also used by [5] for tennis scene classification and segmentation. An HMM was used to fuse audio and visual information. They also used HMMs to model tennis syntax and the hierarchical structure of a tennis match.

In this paper we propose a multistage decision making system that is founded on the concept of visual cues — pieces of visual evidence, characteristic of certain categories of sports that are extracted from key frames. The main decision-making mechanism is a decision tree which generates hypotheses concerning the semantics of the sports video content. The final stage of the decision making process is an HMM system which bridges the gap between the semantic content categorisation defined by the user and the actual visual content categories. Two setups of HMM are considered. In one setup, we constructed and trained an HMM for each sport investigated in our research. This setup is motivated by the fact that each HMM corresponding to a certain sport is constructed and trained independently. However, using this setup, a post-processing is needed to combine the outcomes of the individual HMMs to reach a final decision. In the other setup, a single HMM is constructed with each node representing one of the sports investigated. This setup has the advantage that it requires no post-processing. To reach a decision using this setup, we need to find the single best state sequence for the given observation sequence.

The paper is organised as follows. In Section 2 we give an overview of the system. We briefly describe the cue concept and how to generate cues deemed indicative of sport types in Section 3. Section 4 describes the process of generating sports video content hypotheses using decision trees. The post-processing of the decision tree outputs using two setups of HMMs is discussed in Section 5. The results of experiments designed to test and compare the two HMM setups is presented in Section 6. The paper is concluded in Section 7.

2 System Overview

In this section we give an overview of the system (Figure 2) and describe its various elements. Our goal in this paper is, given a video stream that contains sports material from one or more disciplines, to automatically segment the stream into sequences and label each sequence with the corresponding sport label.

First, the video stream is segmented to shots which are the basic temporal units in our system. For each shot, a number of key frames are extracted. The first stage of the decision-making process is the cue detection. Cue detectors operate on the key frames and generate judgement about the presence or the absence of the objects they try to detect. The shot after this stage is represented by the cues. This is distinct from the conventional approaches which are based on low level generic image features derived from colour and texture. Cues offer higher level representation which is application domain specific. Most importantly, they transform diverse input data structures into a standard form which facilitates the decision making process and promotes modularity (i.e. exploiting additional cues). Examples of cues could be: grass, sky, swimming pool lanes.

The second stage classifies each shot to one of the characteristic views, defined for each sport, using the information provided by the cue detectors. The functionality of this stage is realised by a decision tree classifier. The knowledge embodied in the decision tree is learnt from a set of labelled training samples covering all views the system is trying to detect. The decision tree is then used to classify each shot into one of the sport view categories.

The output of the decision tree may be subject to error due to either errors in the cue extraction or genuine ambiguity, i.e. the presence of cues that are characteristic of more than one discipline (e.g. crowd views). The third stage is designed to minimise this error by exploiting the temporal context using HMMs. HMMs, which process the sequence generated by the decision tree, bridge the gap between the semantic video content labelling by human observer and the data-driven hypotheses generated by automatic classification methods.

The individual stages of the system are described next in more detail.

3 Cue Detectors

In much of the previous work in automatic annotation of video material, the annotation consisted of the output of various feature detectors (i.e. MPEG7 descriptors). By itself, this information bears no semantic connection to the actual

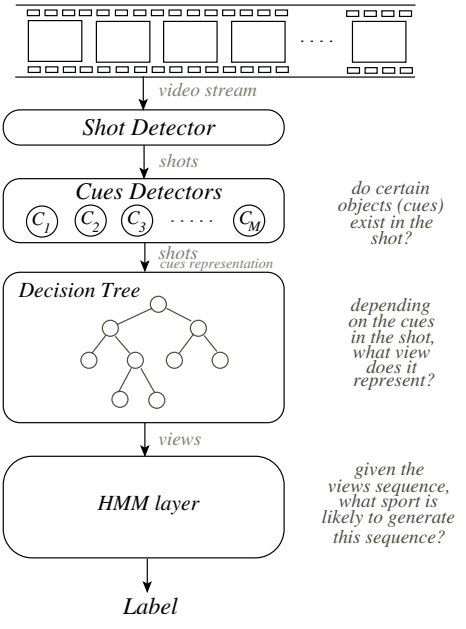


Fig. 2. Proposed System

scene content — it is simply the output of some image processing algorithm. The cue detection approach [6] is taking the process one stage further. In this approach, the connection between low-level image data outputs and the semantics of the scene content can be defined by means of a set of training processes. Thus for example the system can be trained to associate the output of a texture feature detector with crowds of people in the scene. This mechanism can then be used to generate confidence values for the presence of a crowd cue in a scene, based on the scene texture. Different cues can then be combined to generate higher-level information, e.g. the type of sport being played. Figure 3 illustrates the cue generation process which involves three phases. Different cue detection methods have been developed [9, 8, 7]. Each method can be used to form a number of different cue-detectors provided that suitable training data is available.

Let us suppose that we have a set of M trained cue-detectors. Each cue detector operates on the key frame images and generates two pdf values $p(x|C)$ and $p(x|\bar{C})$, where C is the cue looked for by the cue detector and x denotes the measurement vector used. Assuming equal prior probabilities, we can estimate the a posterior probability $P(C|x)$ of an instance of a cue, C, existing in the image as follow:

$$P(C|x) = \frac{p(x|C)}{p(x|C) + p(x|\bar{C})} \tag{1}$$

Thus, for each key frame, we will obtain k values, one for each cue. A shot can then be represented by a vector $S = (\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}j, ..., \mathcal{C}_M)$ where $\mathcal{C}_j$ is the mean

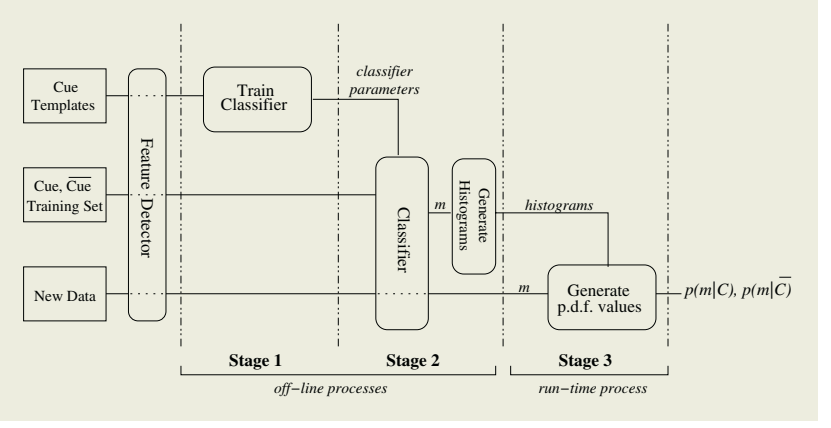


Fig. 3. Creating Cue Evidence

value of the posterior probabilities computed by the j^{th} cue-detector on the key frames that belong to the shot.

4 View Classification

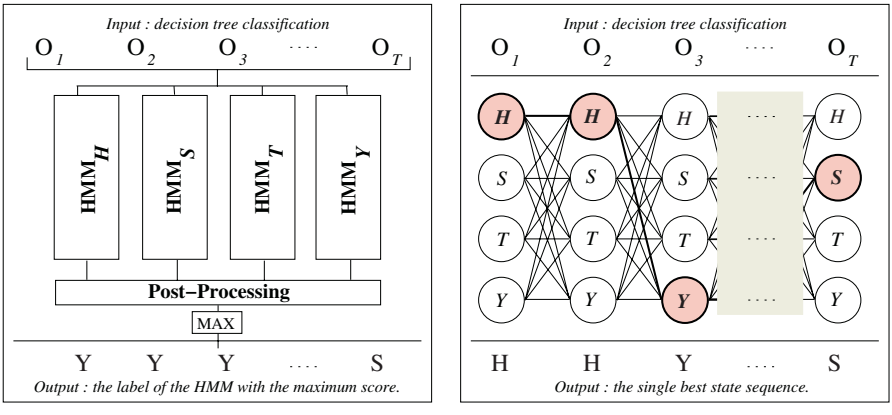
Based on earlier experience [4], we opted for a decision tree learning algorithm to build the model for solving the problem of classifying a shot to one of a set of predefined sport views. The "C4.5" algorithm [10] was adopted. The process of constructing a decision tree classifier requires a set of training examples. Each example is represented by the cue vector, S. A class label, which is typically a camera view, is attached to the examples. A splitting criterion, Information impurity, is used to recursively partition the training set in a way that increases the homogeneity of its partitions. The partitioning stops when one of the stopping rules is triggered at a node. This node becomes a class node and a label which represents the sport view with the largest number of shots is attached to it.

The classification of a shot using decision trees proceeds from top to bottom. Depending on its cue values, S navigates through the decision tree till it reaches a class node. The navigation is guided by the rules of the decision nodes visited. The shot is assigned the label attached to the class node at which its navigation terminates.

5 Post-processing Using HMMs

The HMM (described in detail in [11]) is a powerful tool, widely used in pattern recognition. In this paper, HMM is employed to minimise the ambiguity of classifying using a decision tree by exploiting the temporal context. HMMs bridge the gap between the semantic video content labelling by human observer and the data-driven hypotheses generated by automatic classification methods.

Two HMM setups are considered (Figure 4). In the first setup, we construct and train a separate HMM model for each sport we want to detect. In this setup, the forward algorithm can be used to compute the likelihood that an observation sequence is emitted by an HMM. Figure 5(a) shows the output of four HMMs operating on a sequence. Note that the output of HMM corresponding to the sport of the subsequence will exhibit the least change of all of the HMMs. To exploit this, we compute the discrete derivative of each HMM output and smooth the result with a Gaussian kernel (Figure 5(b)). The subsequence is labelled with the identity of this HMM.



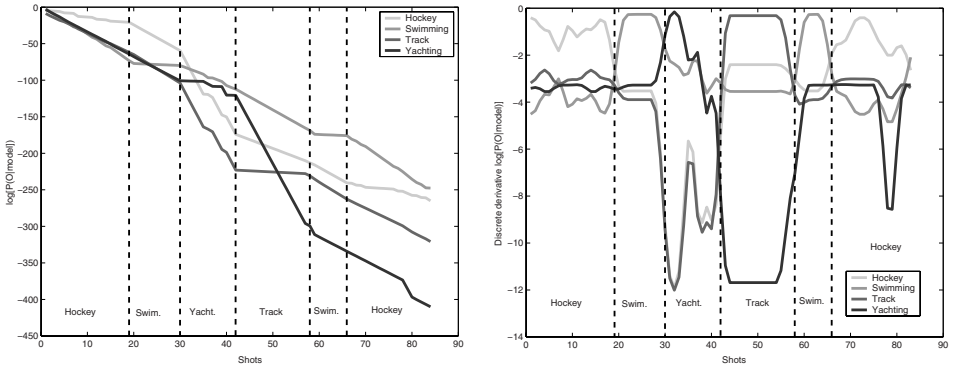
(a) Setup 1 : four competing HMMs (b) Setup 2 : single HMM

Fig. 4. The two HMM setups considered in our system

In the second setup, we construct a single HMM with each node in this HMM representing one of the sports we want to detect. In this setup, the problem we need to solve is to find the single best state sequence given the observation sequence generated by classifying a sequence of shots using the decision tree classifier. The Viterbi algorithm is used to realise the most likely sequence state.

6 Experimental Results

In this section, we describe the experiments to evaluate the proposed system, and compare the results obtained from applying the two HMM setups. The experimental data is taken from video material from the 1992 Barcelona Olympic Games. The material includes four Olympic sport disciplines (hockey (H), swimming (S), track_events (T), yachting (Y)). The material was manually ground-truthed and split into three sets. One set was reserved for training the decision tree classifier; the second set was used for HMM training, and the remaining set was reserved for testing the system. Thirty-seven visual cues were identified, trained and used to generate cue evidence for the study.



(a) Output of the four competing HMM (b) After post-processing

Fig. 5. Output from four HMMs operating on a sequence generated by the decision tree classifier

Table 1. Confusion matrix for sports shot classification of the proposed system using four competing HMMs (Setup 1)

	H	S	T	Y	Recall	Precision
H	**522**	23	30	3	90.3%	87.0%
S	30	**909**	11	14	94.3%	93.4%
T	49	11	**469**	0	88.7%	92.0%
Y	0	30	0	**370**	92.5%	95.6%

Table 2. Confusion matrix for sports shot classification of the proposed system using a single HMM (Setup 2)

	H	S	T	Y	Recall	Precision
H	**551**	5	16	6	95.3%	91.4%
S	46	**903**	14	1	93.7%	98.8%
T	6	6	**516**	1	97.5%	94.5%
Y	0	0	0	**400**	100.0%	98.0%

We tested the proposed system on 52 sequences. Table 1, shows the results obtained from the experiments using Setup 1 in which four competing HMMs are used, one for each discipline. The results obtained from experimenting with the proposed system using single HMM, with each state in this HMM represent one of the sports investigated in this paper, are summarised in Table 2.

The proposed system performed well with both setups. The overall recognition rate is 91.87%(standard deviation = 5.08%) for Setup 1 compared to 95.91%(standard deviation = 3.18%) for Setup 2. We performed a t test and the test suggested that the difference between the performance of the two setups was significant statistically. Moreover, using setup 2 proved to be more convenient since it requires no post-processing on the obtained results. As far as the

accuracy of classification for individual sports is concerned, we noticed that the hockey, track_events and yachting classification rate using Setup 2 were significantly better than when using Setup 1. Swimming performance, was almost the same in both setups.

One advantage of the proposed system is its ability to segment a sequence comprising more than one discipline and label the subsequences with the corresponding sport label. This information can be used to perform further analysis on any subsequence using specialised model once we know its coarse label.

Doing more analysis on the results, we noticed that just over 70% of the subsequences were correctly labelled and 3% were mislabelled. The boundaries of the remaining 27% of the subsequences, do not exactly correspond to the groundtruthed test data. However, in 59% of the latter cases, the errors in the boundaries are due to ambiguity in the material rather than the classification system, i.e. crowd shot at the beginning or the end of a sport event.

7 Conclusion and Future Work

In this paper, a multi-stage decision-making system for sports video classification was proposed. The first stage of the decision-making process detects application-specific cues. The second stage attaches a label, from a set of prototypical views of each sport, to each shot, using the information provided by the cue detection stage. The functionality of this stage is realised by a decision tree classifier. The third stage uses HMMs to process the sequence of view labels generated by the decision tree. The output of this stage is a final decision regarding the identity of the sport represented by the sequence, taking advantage of the temporal context. We experimented with the proposed system using two setups of HMM, four competing HMMs, one for each discipline, and a single HMM with each node representing a sport. It was noticed from the experiments that using single HMM delivered better results.

Our future plans include providing cues that deal with modalities other than the visual ones. They are expected not only to improve the sports video categorisation performance but also help to detect highlights such as hockey goal, etc. It is intended to use audio, speech and motion cues for this purpose.

Acknowledgements

This work was supported by the IST-2001-34401 VAMPIRE project funded by the European IST Programme.

References

1. P. Chang, M. Han, and Y. Gong. Extract Highlights From Baseball Game Video With Hidden Markov Models. In *IEEE International Conference on Image Processing (ICIP'02)*, 2002.

2. A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic Soccer Video Analysis and Summarization. *IEEE Transactions on Image Processing*, 12(8):796–807, July 2003.
3. Y. Gong, T.S Lim, and H.C. Chua. Automatic Parsing of TV Soccer Programs. In *IEEE International Conference on Multimedia Computing and Systems*, pages 167 – 174, May 1995.
4. E. Jaser, J. Kittler, and W. Christmas. Building Classifier Ensembles for Automatic Sports Classification. In Roli F Windeatt T, editor, *Proceedings of the 4th International Workshop on Multiple Clasifier Systems (MCS 2003)*, volume 2709 of *Lecture Notes in Computer Science*, pages 366–374. Springer-Verlag, June 2003.
5. E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. HMM Based Structuring of Tennis Videos Using Visual and Audio Cues. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 309–312, July 2003.
6. J. Kittler, K. Messer, W. Christmas, B Levienaise-Obadia, and D. Koubaroulis. Generation of Semantic Cues for Sports Video Annotation. In *Proceedings of the 2001 International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece*, pages 26–29, October 2001.
7. B. Levienaise-Obadia, J. Kittler, and W. Christmas. Defining Quantisation Strategies and a Perceptual Similarity Measure for Texture-Based Aannotation and Retrieval. In *In IEEE, editor, ICPR'2000*, volume III, 2000.
8. J. Matas, D. Koubaroulis, and J. Kittler. Colour Image Retrieval and Object Recognition Using the Multimodal Neighbourhood Signature. In *D Vernon, editor, Proceedings of the European Conference on Computer Vision LNCS*, volume 1842, pages 48–64, 2000.
9. K. Messer and J. Kittler. A Region-Based Image Database System Using Colour and Texture. In *Pattern Recognition Letters*, page 1323 1330, 1999.
10. J. R. Quinlan. *C4.5 : Programs for machine learning*. Morgan Kaufmann, 1993.
11. Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE*, 77(2):257–286, 1989.
12. P. Xu, L. Xie, S. Chang, A. Divakaram, A. Vetro, and S. Sun. Algorithms and System for Segmentation and Structure Analysis in Soccer Video. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2001.