論文 / 著書情報
Article / Book Information

Title(English)	Robust Scene Recognition for Baseball Broadcast
Authors(English)	Koichi Shinoda, Sadaoki Furui
Citation(English)	Proc. International Symposium on Large-Scale Knowledge Resources (LKR), Vol. , No. , pp. 91-94
発行日 / Pub. date	2006, 3

Proc. International Symposium on Large-scale Knowledge Resources (LKR2006), 2006.3

Robust Scene Recognition for Baseball Broadcast

Koichi Shinoda and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technolgoy, Japan
shinoda@cs.titech.ac.jp

Abstract

This paper introduces a statistical framework for recognizing scenes from a baseball broadcast video. Inspired by the successes of statistical approaches in speech recognition field, we propose a data-driven approach to provide robust scene recognition. We use several global features and apply multi-stream Hidden Markov Models (HMMs) to control the weights among them. To achieve robustness against new scenes, we use a common simple structure for all the HMMs. In addition, game adaptation is applied to achieve greater robustness against differences in environmental conditions among games.

1. Introduction

Recent advances in computer technology, particularly in storage technology, have resulted in significant increases in the number and quality of video databases. Accordingly, it has become difficult for ordinary people to browse the whole content of each video database. An index describing its content is strongly required for searching and summarization. While the construction of such an index is mostly carried out by some experts who manually assign a limited number of keywords to the video content, the specialist nature of this work makes it an expensive and time-consuming task. Therefore, automatic indexing using pattern recognition techniques for video contents, which we call Content-Based Video Information Retrieval (CBVIR), has been studied extensively [1].

Compared with other data sources, broadcast data and sports broadcasts in particular have well-defined structures and users' demands are relatively clear. This has made scene recognition in sports broadcast data a topic for many researches in recent years [2, 3]. In this paper, we focus on scene recognition in baseball broadcast data.

In recent years, many methods for recognizing scenes from baseball broadcasts have been proposed. In a baseball broadcast, the minimum unit is a *frame*, a static image. Multiple frames recorded by a single fixed camera form a *shot*. A sequence of these shots forms a *scene*. Scenes that have significant information for understanding games are *highlights*. The contexts or transitions between those shots provide useful information for scene extraction. For example, in Chang *et al.*'s work [2], video data is first segmented into shots. Then, extraction is applied to these shot sequences based on Hidden Markov Models (HMMs) in which each state represents a *shot type*. Li *et al.* [3] also proposed an HMM-based framework to distinguish *play* and *non-play* scenes.

In these works, domain-specific knowledge about shot types and the transitions among them were used intensively to improve the system's performance. But the large variety of transitions among shots from game to game and the difficulty of classifying shot types are still problems for these approaches. Systems resulting from these studies may not be robust enough to apply to real applications in general.

Inspired by the successes of using statistical frameworks in the speech recognition field (e.g., [4]), we proposed a *data-driven* approach to provide a *robust* scene recognition [5]. In our approach, we regarded a shot as being analogous to a *phone*, and a scene to a *word* and utilized the framework of *continuous speech recognition* in scene recognition. Given a sequence of observed feature vectors $O = o_1, \ldots, o_m$ (m is the number of frames), the probability of scene sequence $H = h_1, \ldots, h_n$ is:

$$P(H|O) \propto P(O|H)P(H), \tag{1}$$

where $P(H)$ indicates the probability of the sequence H (*language model*) and $P(O|H)$ is the probability of O being observed in the scene sequence H (*video model*). The sequence H that maximizes $P(H|O)$ is the recognition result. In this paper, we focus on modeling the *video model* $P(O|H)$. The model we use here is a multi-stream HMM, which can control the weights among different features. To achieve robustness against unknown scenes, a relatively simple structure is used for all scene models instead of a different structure for each model as in the previous studies. In addition, an *game adaptation* method is applied to adjust the model dispersion among games.

2. Feature extraction from video data

In order to make our framework generally applicable, we avoid using any game-specific features, such as those related to infield color or uniform color. Moreover, we use only global features and do not use features related to specific objects because it is not alway easy to extract these objects from video in various conditions. We investigate the following four features in this study [5].

Principal component features (PFs) By decreasing the dimensions by using Principal Component Analysis (PCA), we expect to remove noises unrelated to scene characteristics and select only scene-relevant global features [6]. We apply PCA to a gray-scale image of each frame and use the first 60 eigenvectors.

Fractal features (FFs) Fractal features, which consist of a *fractal dimension* and a *complexity*, represent global information about textures of the still image of each frame [7]. We calculate these two features in each of 2x3 sub-blocks to create a 12-dimension vector for each frame.

Difference features (DFs) While PFs and FFs are expected to be sufficient for representing global information for a still image, such representation for video images requires additional information for describing objects moving in

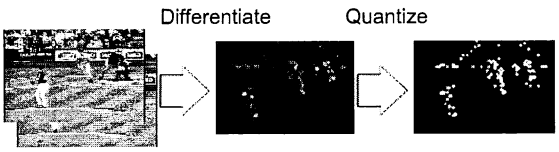

Figure 1: Calculation of difference features (DFs).

a video stream, information that is strongly related to scene characteristics. Here, we simply use differences between successive two frames [8](Fig. 1). We calculate the distribution of moving parts (white pixels in Fig. 1) and their ratio to the whole image.

Camera-motion features (CFs) Camera motions in a baseball broadcast consist of three kind of movements: pan, tilt, and zoom. Camera shots of the same shot type tend to have similar camera motions. In this study, we use pan and tilt features.

3. Scene modeling by multi-stream HMMs

HMMs are effective models for time-varying patterns and have been used widely to model scenes of sports video [2, 9]. In the speech recognition field, together with HMMs, *multi-stream HMMs*, in which features are split into separate streams, have been widely used. By using multi-stream HMMs, we can control the weights among different types of features in an optimization process. Conventional methods of recognizing scenes based on HMMs have not yet exploited the usefulness of this model.

In conventional HMM-based scene recognition methods (e.g., [2]), each state of an HMM is usually assigned to a specific shot type, and the HMM of each scene has a specific topology that is determined heuristically. Inspired by the effectiveness of the data-driven approach used in speech recognition, we do not explicitly define a specific topology for each scene, but use a common left-to-right HMM for all scenes.The reason for this is that, in real data, while the shot transition of each scene varies greatly, few clues about the underlying shot transition are apparent. Using this data-driven approach makes it easy to prepare scene models and to achieve robustness against unknown data. Our framework can be applied without any modification to recognize new scene types or when the amount of available training data increases.

In multi-stream HMMs, each state j has an associated observation probability distribution $b_j(\mathbf{o}_t)$ which determines the probability of generating observation $\mathbf{o}_t$ at time t, and each pair of states i and j has an associated transition probability a_{ij}. The output probability $b_j(\mathbf{o}_t)$ of state j is calculated by multiplying the probability of each output stream s by its respective weight w_s:

$$b_j(\mathbf{o}_t) = \prod_{s=1}^{S} b_{js}(\mathbf{o}_{ts})^{w_s}, \qquad \sum_{s=1}^{S} w_s = 1, \tag{2}$$

where $b_{js}(\mathbf{o}_{ts})$ denotes the probability of the s-th output stream at state j, and S denotes the total number of streams.

Let $\mathbf{O}_k = (\mathbf{o}_{k1}, \mathbf{o}_{k2}, \mathbf{o}_{k3}, \mathbf{o}_{k4})^T$ be the feature vector containing the four features calculated from the k-th frame. For each scene model H, HMMs parameter a_{ij} and b_j are learned from training data using the *Baum-Welch algorithm*. A single

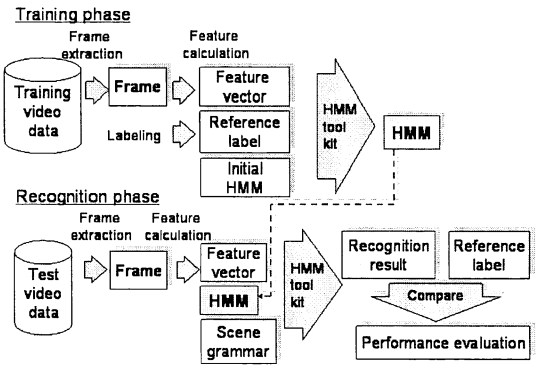

Figure 2: Scene model training and continuous scene recognition

Gaussian distribution is used as the output probability. Given a feature vector sequence $\mathbf{O}_1\mathbf{O}_2 \ldots \mathbf{O}_m$, the corresponding scene sequence $H_1H_2 \ldots H_n$ is recognized by the *Viterbi algorithm*. In the recognition process, data is matched against a network of HMMs defined by a *grammar* (representing a *language model*), which expresses the rule of possible scene-label sequences in the data [4, 10].

4. Statistical scene recognition

We investigated a scene recognition method using multi-stream HMMs from baseball video data, which contains two phases: (Fig. 2):

1. In the training phase, frames are extracted from video data, and features are calculated from each frame to form an individual feature vector. One HMM is prepared for each scene, and its parameters are estimated on the basis of a training set of feature vectors and on reference labels with boundary information that has been prepared manually.
2. In the recognition phase, we extract frames from video data and calculate feature vectors in exactly the same way as in the training phase. Then, using the trained HMMs, we conduct scene recognition on the test data. Here, we use the simple *grammar* represented by the Chomsky expression $< H_1|H_2|...|H_n >$ to build the scene HMM network used in the recognition process. The grammar indicates that the scene sequence in data is a combination of many scenes in an arbitrary order. An arbitrary scene is extracted by using time information associated with recognition results.

The output is a sequence of scene names with time boundary information. This method has the merit of being capable of recognizing scenes directly without information about scene boundaries or shot types, so recognition performance is not influenced by the accuracy of the shot classification process.

5. Game adaptation

The parameters for scene HMMs are estimated using data from many games. This model is a *game-independent* (GI) model

Table 1: F-measures (%) for the scene recognition

Stream weights (PF, FF, DF, CF)	Game Adaptation	F-measure(%)
(1.00, 0.00, 0.00, 0.00)	No	61.2
(0.00, 1.00, 0.00, 0.00)	No	41.4
(0.00, 0.00, 1.00, 0.00)	No	**64.1**
(0.00, 0.00, 0.00, 1.00)	No	22.1
(0.25, 0.25, 0.25, 0.25)	No	64.3
(0.08, 0.00, 0.92, 0.00)	No	**77.4**
(0.08, 0.00, 0.92, 0.00)	Yes	**81.1**

that represents average characteristics that appear in all the games. In speech recognition, by using so-called *speaker adaptation*, a *speaker-independent* model can be adjusted to a specific speaker to improve recognition rate given a small amount of data for that speaker. The adjusted model is called a *speaker-dependent* model. We can apply the same technique to our framework to adjust a model to make it a *game-dependent* (GD) model. We use an adaptation method using Maximum Likelihood Linear Regression (MLLR) [11]. MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically, MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of Gaussian mixture HMMs. The variance parameters are not adapted in our method since adapting them is less effective than adapting the mean.

In MLLR adaptation, the mean parameter μ in each state of the HMMs is transformed as follows:

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b}, \tag{3}$$

where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{b}$ is an n-dimension vector, which are obtained by solving a maximization problem using the *Expectation Maximization* (EM) algorithm. We use the GI model as an initial model for the adaptation.

The effect of this transformation is to transform the means in the HMM so that each state in the HMMs is more likely to generate the adaptation data. If the adaptation data is labeled we have a *supervised adaptation*. Otherwise, *unsupervised adaptation* is used in which the recognition results are used as supervising signals for the adaptation process. Since scene labels are not available in a real situation, we use unsupervised adaptation.

6. Experiments

6.1. Experimental conditions

We used 40 hours of baseball broadcast video provided by NHK (Japan Broadcasting Corporation) Lab, which consisted of ten games. From this full data, 4.5 hours of digest data were created for tractability by removing replays, CG effects, and other scenes such as regular strikes and balls, that would not attract much interest from users. To this data, we applied the 8 scene labels such as timely hit and strike out. Four hours of the data, consisting of nine games, were used for training and the other 30 minutes, consisting of one game, for testing. For the evaluation, we used *F-measure*, a harmonic average of Precision and Recall.

6.2. Results

First, we evaluated the scene recognition performance of our proposed method. One scene HMM was prepared for each scene label. Each scene HMM had the same number of states, 75, which was optimized in our preliminary experiment. The features used in this experiment were the four features explained in Section 2: principal component features (PFs), fractal features (FFs), difference features (DFs), and camera-motion features (CFs). The weights among these features in the multi-stream HMMs were optimized by our preliminary experiments using test data in this study.

The results when the weights for the four streams were varied in several ways are shown in Table 1. While the F-measure was 64.3% when the stream weights were equal among the four streams, (0.25, 0.25, 0.25, 0.25), it was 77.4% when the stream weights were (0.08, 0.0, 0.92, 0.0). The F-measure was improved by 13.1 points. This result confirmed the effectiveness of our multi-stream HMMs. It also indicated that PFs and DFs are more effective than the other two features.

Then, we applied the game adaptation using MLLR to the multi-stream HMM with the stream weights (0.08, 0.0, 0.92, 0.0), which had the highest performance in the previous experiments. The result are also shown in Table 1. The result after unsupervised adaptation was 81.1% with improvements of 3.7 points. This result confirmed that the game adaptation was effective for scene recognition.

7. Discussion

Recently CBVIR for sports videos has been extensively studied. Their targets include baseball [2, 12], soccer [13, 14], tennis [15], basketball [16], and American football [17]. While all of those studies employed shot boundary information, our proposed method, on the other hand, uses no shot boundary information. Thus, our system is more robust than the conventional methods because the shot boundaries are often difficult to detect in real applications.

Most conventional methods used local features that are dependent on the target sport. Some examples are "grass and sand feature" in baseball [2] and "line marks" in soccer [13, 14]. Our method, on the other hand, employs only global features, such as PFs and DFs that do not depend on the kind of target sport. Therefore, it can be applied to other sports with little effort. Table 1 shows that the four features, PFs, FFs, DFs, and CFs, have different effects on the performance of scene recognition. When they were combined, PFs and DFs proved to be especially important. When they were used one by one, their F-measures were 61.2%, 41.4%, 64.4%, and 22.1%, respectively. The difference features (DFs) gave the highest value. This result indicates that the importance of *dynamic* features is equal to or greater than that of *static* features, though most conventional methods used only static features. Camera-motion features (CFs), on the other hand, gave the lowest performance among the four features. Only pan and tilt features may be insufficient to represent the characteristics of the camera movement for each scene.

In this work, we employed only features in the video stream and did not use features in other modes such as speech or closed captions. Some previous studies [17, 12] showed that multimodal approach involving those other streams improved the recognition performance. It is easy to include other mode in our approach based on multi-stream HMMs. Research in this direction is promising.

We proposed a novel game adaptation scheme in Section 5 and confirmed its effectiveness as shown in Table 1. Our system adapts to the differences among games without any additional information and achieves more stable performance. This scheme can be easily applied to other statistical approaches using HMMs. While unsupervised adaptation was used in this research, supervised adaptation should be used when data for adaptation is available. For example, let us recognize a game between Team A and Team B in Stadium C. In a real situation, it is likely that these two teams have had many games in the same stadium in the past. The data of those past games can be used for the supervised adaptation to improve more the performance of scene recognition. Since the numbers of teams and stadiums are limited, it is feasible to prepare such adaptation data.

8. Conclusions

A scene recognition framework for baseball broadcasts was proposed. In this framework, multi-stream HMMs are used to model each scene. Since the structure of these models is simple and identical for all the scenes, it is robust against variations in data and against the unbalanced occurrences of scenes. Four features, principal component features (PFs), fractal features (FFs), difference features (DFs), and camera-motion features (CFs) are used in this method. This method was evaluated by using digest data of baseball broadcasts. The F-measure was improved by 12.9 points when multi-stream HMMs were used and was further improved by 3.7 point when game adaptation was applied. The final F-measure achieved by our method was 81.1%.

This study should be regarded as the starting point of video indexing using the data-driven approach which is often used in speech recognition. Many problems still remain to be solved. First, while the proposed method was evaluated only for digest data in this study, an evaluation with whole-game data should be carried out. In this case, it is important to build a *language model* ($P(H)$ in Eq. (1)) that represents the scene contexts and to deal with the *out-of-vocabulary* scenes, which do not appear in training data. Second, there are many other features that should be effective for our method. One example is color information. Such features need to be included in our framework. Furthermore, the stream weights for the multi-stream HMMs were optimized by using test data in this study. Weight optimization methods using discriminative training should be investigated. Finally, we plan to extend our framework using multi-stream HMMs to a *multi-modal* recognition framework that deals not only with video mode but also with other modes such as speech and text.

9. Acknowledgments

We thank NHK Lab. for permission to use its baseball broadcast video database and for valuable co-operation in the collaborative research.

10. References

[1] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *Journal of Visual Computation and Image Representation*, vol. 10, no. 2, pp. 78–112, 1999.

[2] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," *Proc. of the Int. Conf. on Image Processing (ICIP '02)*, vol. 1, pp. I–609–612, 2002.

[3] B. Li and M. I. Sezan, "Event detection and summarization in sports video," *Content-Based Access of Image and Video Libraries 2001 (CBAIVL 2001)*, vol. 9, no. 8, pp. 132–138, 2001.

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

[5] H.-B. Nguyen, K.Shinoda, and S. Furui, "Robust highlight extraction using multi-stream hidden markov models for baseball video," *Proc. International Conference on Image Processing 2005 (ICIP2005)*, vol. 3, pp. 173–176, 2005.

[6] E. Sahouria and A. Zakhor, "Content analysis of video using principal component," *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 1290–1298, 1999.

[7] T. Mochizuki, M. Fujii, and T. Ito, "Image retrieval using a new fractal feature and robust structural information," *The Journal of the Institute of Image Information and Television Engineers*, vol. 57, no. 6, pp. 719–728, 2003, (in Japanese).

[8] S. Eickeler and S. Müller, "Content-based video indexing of TV broadcast news using hidden Markov models," *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '99)*, pp. 2997–3000, 1999.

[9] D. Zhong and S.F. Chang, "Structure analysis of sports video using domain models," *Proc. of Int. Conf. on Multimedia and Expo*, pp. 920–923, 2001.

[10] S. Young, *The HTK book (for HTK version 3.2)*, http://htk.eng.cam.ac.uk/docs/docs.shtml, 2002.

[11] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, , no. 2, pp. 171–185, 1995.

[12] Y. Gong, M. Han, W. Hua, and W. Xu, "Maximum entropy model-based baseball highlight detection and classification," *International Journal of Computer Vision and Image Understanding*, vol. 96, pp. 181–199, 2004.

[13] Y. Gong, L.-T. Sin, C.-H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of tv soccer programs," *Proc. International Conference on Multimedia Computing and Systems*, pp. 167–174, 1995.

[14] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 11, no. 2, pp. 135–145, 2003.

[15] E. Kijak, L. Oisel, and P. Gros, "Hierarchical structure analysis of sport videos using hmms," *Proc. ICIP 2003*, vol. 3, pp. 1025–1028, 2003.

[16] G. Xu, Y.-F. Ma, H.-J. Zhang, and S. Yang, "Motion based event recognition using hmm," *Proc. ICPR 2002*, vol. 2, pp. 831–834, 2002.

[17] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 68–75, 2002.