Statistical Framework for Shot Segmentation and Classification in Sports Video*

Ying Yang[1,2], Shouxun Lin[1], Yongdong Zhang[1], and Sheng Tang[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100085, China
{yyang, sxlin, zhyd, ts}@ict.ac.cn

Abstract. In this paper, a novel statistical framework is proposed for shot segmentation and classification. The proposed framework segments and classifies shots simultaneously using same difference features based on statistical inference. The task of shot segmentation and classification is taken as finding the most possible shot sequence given feature sequences, and it can be formulated by a conditional probability which can be divided into a shot sequence probability and a feature sequence probability. Shot sequence probability is derived from relations between adjacent shots by Bi-gram, and feature sequence probability is dependent on inherent character of shot modeled by HMM. Thus, the proposed framework segments shot considering the character of intra-shot to classify shot, while classifies shot considering character of inter-shot to segment shot, which obtain more accurate results. Experimental results on soccer and badminton videos are promising, and demonstrate the effectiveness of the proposed framework.

Keywords: shot, segmentation, classification, statistical framework.

1 Introduction

In recent years, there has been increasing research interests in sports video analysis due to its tremendous commercial potentials, such as sports video indexing, retrieval and abstraction. Sports video can be decomposed into several types of video shots which are sequences of frames taken contiguously by a single camera. In sports video analysis, shot segmentation and classification play an important role for shots are often basic processing unit and give potential semantic hints.

Much work was done on shot segmentation and classification, most of which take shot segmentation and classification as a two-stage successive process, and perform the two stages independently using different features without considering the relationship between them. Many shot segmentation algorithms have

* This research was supported by National Basic Research Program of China (973 Program, 2007CB311100), Beijing Science and Technology Planning Program of China (D0106008040291), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416).

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 106–115, 2007.

been developed by measuring the similarity of adjacent shots [1]. However, these algorithms get poorer performance on sports video due to frequent panning and zooming caused by fast camera motion. After shot segmentation, shot classification is performed on each segment for higher level video content analysis. Some work classified sports video shots based on domain rules of certain sports game [2,3], such as classifying soccer video shots into long shots, medium shots and close-up shots using dominant color. Others tried a unified method to solve this problem using SVM or HMM [4,5,6,7,8]. For these approaches, since classification is done after segmentation independently, the incorrect segmentation will have a negative effect on shot classification.

In this paper, a novel statistical framework is proposed for shot segmentation and classification. Compared with previous work, the proposed framework classifies and segments shot simultaneously using same difference features based on statistical inferences. The task of shot segmentation and classification is taken as finding the most possible shot sequence given feature sequence, and it can be formulated by a conditional probability which can be divided into a shot sequence probability and a feature sequence probability. Shot sequence probability is derived from relations between adjacent shots models modeled by Bi-gram, and feature sequence probability is dependent on inherent characters of shot modeled by HMM(Hidden Markov Model). Therefore, the proposed framework segments shot considering characters of intra-shot, while classifies shot considering characters of inter-shot, which is a global search on all the possible shot sequence for the best shot sequence matching feature sequence.

The rest of paper is organized as follows. In Section2, the main idea of the statistical framework is presented. Section 3 gives the details of shot segmentation and classification based on the proposed statistical framework. To evaluate the performance of this framework, two applications in soccer and badminton videos and result analysis are described in Section4. Finally, conclusions are drawn and future work is discussed in Section 5.

2 Main Idea of the Statistical Framework

Suppose that a video stream is composed of a sequence of shots, denoted by $H = h_1 h_2 \ldots h_t$. After feature extraction, the video stream can be viewed and manipulated as a sequence of feature vectors, denoted by $O = o_1 o_2 \ldots o_T$. Hence, the task of shot segmentation and classification can be seen as mapping the sequence of feature vectors to a sequence of shots, and the best mapping is the expected result of shot segmentation and classification. Hence, in the proposed framework, the task is interpreted as finding a shot sequence that maximize the conditional probability of H under condition O, namely, finding

$$\hat{H} = \arg\max_H \{\mathrm{P}(H|O)\} = \arg\max_H \{\mathrm{P}(H) \cdot \mathrm{P}(O|H)/\mathrm{P}(O)\} \quad (1)$$

The equation is transformed by applying Bayes' theorem. Since $\mathrm{P}(O)$ is constant for O is a known sequence, the problem can be simplified as following

$$\hat{H} = \arg\max_H \{\mathrm{P}(H) \cdot \mathrm{P}(O|H)\} \quad (2)$$

The calculation of above probability involves two types of probability distribution, i.e. $\mathrm{P}(H)$ and $\mathrm{P}(O|H)$.The former indicates probability of shot sequence without effect of features, and the latter indicates probability of feature sequence under a given shot sequence. So we call them *Shot Sequence Probability* (SSP) and *Feature Sequence Probability* (FSP), respectively.

2.1 Shot Sequence Probability

Shot sequence H is composed of successive shots of different categories, so the shot sequence probability is dependent on the transition probability between adjacent shots. Since shot sequence is a temporal sequence, the appearance of the present shot is only related to the appearances of shots prior to it. Therefore, shot sequence can be taken as a Markov process. We suppose the shot sequence is a 1 D Markov, namely, the appearance of present shot is only related to the last shot, which can be formulated by the following equation

$$\mathrm{P}(h_m|h_1h_2\ldots h_{m-1}) = \mathrm{P}(h_m|h_{m-1}) \tag{3}$$

This assumption is reasonable since sports video shots regularly alternate to exhibit certain semantic content according to the play status. Hence, shot sequence probability $\mathrm{P}(H)$ can be calculated by

$$\mathrm{P}(H) = \mathrm{P}(h_1h_2\ldots h_{t-1}h_t) = \mathrm{P}(h_1)\cdot\mathrm{P}(h_2|h_1)\cdot\ldots\mathrm{P}(h_t|h_{t-1}) \tag{4}$$

where $\mathrm{P}(h_i)$ denotes the initial distribution probability of shot h_i , and $\mathrm{P}(h_j|h_i)$ denotes transition probability between shot h_i and h_j , which we called Bi-gram in our framework. So SSP can be deduced according to the Bi-gram.

2.2 Feature Sequence Probability

Feature sequence is also a temporal sequence which indicates a certain shot sequence. Since HMM has been successful used in speech recognition to model the temporal evolution of features in a word [9], we use HMM to model sports video shot for the word and shot have similar temporal structure.

Hence, feature sequence of a shot is an observed sequence of a given shot HMM, and each emitting state of HMM produces a feature vector in the feature sequence [9]. So feature sequence probability $\mathrm{P}(O|H)$ can be transformed into

$$\mathrm{P}(O|H) = \sum_S \mathrm{P}(O,S|H) \tag{5}$$

where $S = s_1s_2\ldots s_T$ is the state sequence which emits feature sequence $O = o_1o_2\ldots o_T$ through the link of all the shot HMMs which we called a super HMM. A super HMM is obtained by concatenating the corresponding shot HMMs using a pronunciation lexicon [10]. So $\mathrm{P}(O,S|H)$ can be derived from

$$\mathrm{P}(O,S|H) = \prod_{t=1}^{T}\mathrm{P}(o_t,s_t|s_{t-1},H) = \prod_{t=1}^{T}[\mathrm{P}(s_t|s_{t-1},H)\cdot\mathrm{P}(o_t|s_t)] \tag{6}$$

$\mathrm{P}(o_t|s)$ is the emission probability distribution of state s, and $\mathrm{P}(s_t|s_{t-1}, H)$ is the transition probability between two states. State transitions of intra-HMM are determined from HMM parameters, and state transitions of inter-HMM are determined by Bi-gram Therefore, these two components can be derived from the parameters of shot HMM and Bi-gram.

3 Shot Segmentation and Classification in Sports Video

In this section, we present the semantic shot segmentation and classification based on the statistical framework presented in Section 2, and shots are classified into three categories including Long Shot (LS), Medium Shot (MS) and Close-up Shot (CS). As described in Section 2, the task of shot segmentation and classification is taken as solving the problem of a maximum conditional probability $\mathrm{P}(H|O)$, which is dependent on SSP and FSP derived from Bi-gram and the parameters of shot HMM. Hence, Bi-gram and shot HMM are the two keys to shot segmentation and classification. The whole framework is shown in Fig.1.

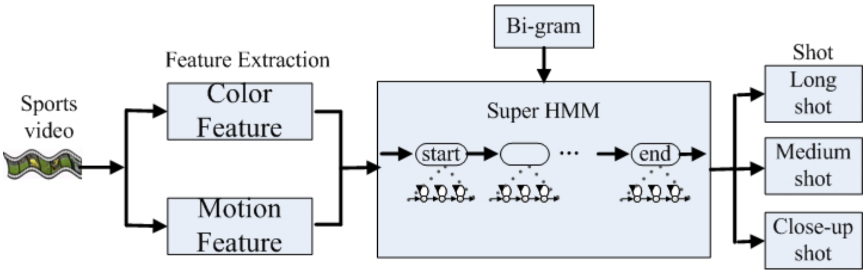

Fig. 1. Framework for shot segmentation and classification

Since parameters of HMM are estimated by EM algorithm using the feature vector sequence [11], appropriate features are vital to the HMM construction and can better explain the temporal evolution in a shot. So in section 3.1, feature extraction is introduced. Section 3.2 and Section 3.3 present the shot HMM and Bi-gram constructions, respectively. In section 3.4, the procedure of simultaneous shot segmentation and classification is discussed.

3.1 Feature Extraction

As mentioned in Section 2, feature sequence of a shot is the observed vector sequence of shot HMM. So each shot is partitioned into segments to extract features from each segment, and features of all the segments form the feature vector sequence. Therefore, shot segment can be one or more consecutive frames, which called as Shot Segment Unit (SSU). Given the length of SSU, the Segmenting Rate (SR) is required at frame level to determine the space of two successive SSUs. To remain more information of shot, the size of SR may be smaller than that of SSU, as shown in Fig.2.

After the magnitudes of SSU and SR are set, feature vector is extracted from each SSU. Two classes of color and motion related features are used in our work for they are generic and can be easily computed, which can be extended to most categories of sports game. Features are firstly extracted from each frame, and then feature values of a SSU are set as the mean of corresponding feature values of all the frames in it.

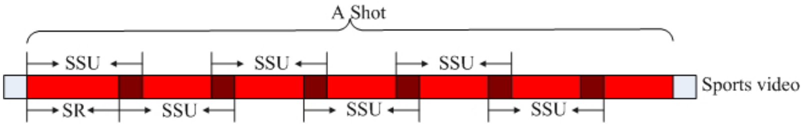

Fig. 2. SSU segmentation

Color Related Features. Since the three categories of shots have different playfield and player sizes, such as LS has the largest playfield view, MS have a smaller playfield and a whole player, frames in each type of shot have a distinct color distribution which is differentiated from that of other types of shots. Hence, color features are derived from L, U and V components for CIE LUV color space is approximately perceptual uniform and computed by the following equations

$$\begin{cases} L_f = \sum_{\text{all pixels in frame } f} L(x,y)/\text{numbers of pixels in frame } f \\ U_f = \sum_{\text{all pixels in frame } f} U(x,y)/\text{numbers of pixels in frame } f \\ V_f = \sum_{\text{all pixels in frame } f} V(x,y)/\text{numbers of pixels in frame } f \end{cases} \tag{7}$$

where L_f , U_f and V_f are the 3 basic color features of a frame f, and $L(x,y)$, $U(x,y)$, $V(x,y)$ are the L, U, V components of pixel (x,y) in f , respectively.

Motion Related Features. Motion is another important factor to distinguish shot for different shots reflect various camera motions, such as LS usually has relatively much smaller motion than CS. 3 basic motion features are extracted for a frame, which are frame difference D_f , compensated frame difference C_f and motion magnitude M_f .

$$D_f = \sum_{i=0}^{255} (H(f,i) - H(f-1,i))^2/\text{number of pixels in frame } f \tag{8}$$

where $H(f,i)$ is the number of pixels of color i in frame f. The other two basic motion features are built on block-based motion analysis. For each block $B_f(s,t)$ at location (s,t) in present frame f, a block $B^*_{f-1}(u^*,v^*)$ from the previous frame $f-1$ is found to best match it. So C_f and M_f are computed as follows, where G_f is the set of all blocks in frame f.

$$\begin{cases} C_f = \sum_{B_f(s,t)\in G_f} (\sum_{i=0}^{255} (H_f(B_f(s,t),i) - H_{f-1}(B^*_{f-1}(u^*,v^*),i))^2) \\ M_f = \sum_{B_f(s,t)\in G_f} ((s-u^*)^2 + (t-v^*)^2)^{1/2} \end{cases} \tag{9}$$

Difference is widely used in signal processing, which can interpret the trend of transformation. Therefore, differences are used to express the trend of variety of color and motion of different types of shots. Given 6 basic features, the 1$^{\text{st}}$ and 2$^{\text{nd}}$ differences of basic features are computed as follows

$$\nabla^1 Fea_f = Fea_f - Fea_{f-1} \qquad \nabla^2 Fea_f = \nabla^1 Fea_f - \nabla^1 Fea_{f-1} \tag{10}$$

where Fea_f denotes one of the 6 basic features, $\nabla^k Fea_f$ $(k = 1, 2)$ denotes the k^{th} difference of basic feature Fea_f of frame f. As a result, there is a 18-D feature vector for each frame and SSU.

3.2 Shot HMM Construction

Since there are 3 categories of shots to be classified, a general solution is to build 3 HMMs modeling 3 types of shots, respectively. However, this method only considers the temporal evolution of SSUs in a single shot, and doesn't take into account the temporal evolution of SSUs at the transitions between adjacent shots. In fact, in sports videos, the alternation of various types of shots exhibits certain rules, such as MS is mainly followed by LS, while LS is often followed by CS. Therefore, to better simulate the temporal evolutions of SSUs of intra-shot and inter-shot, a context-dependent shot model is introduced, which is defined as tri-shot HMM. Tri-shot HMM, just as its name indicates, models the temporal evolution of SSUs in three shot including prior shot, present shot and next shot. Compared with an ordinary shot model, tri-shot model is trained use feature vector sequence of the 3 involved shots. Since we have 3 categories of shots, there are $3^3 = 27$ tri-shots in total.

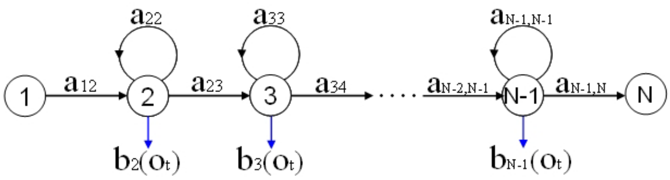

Fig. 3. Topology of HMM structure

For simplicity, we use a left-to-right HMM to represent each tri-shot HMM. The middle states are emitting states with output probability distributions, as shown in Fig.3. Each HMM contains 5 states, and Gaussian Mixture Models are used to express continuous observation densities of the emitting states.

3.3 Bi-gram Construction

As described in Section 2, Bi-gram denotes the transition probability between two adjacent shots, which indicates the possibility of a transit of a shot to another shot. In sports video, various types of shots display different play status and they don't appear randomly. For example, after LS is shown to exhibit the global game status, MS or CS is often shown to track the player or ball. Therefore, Bi-gram can be calculated according to the statistics of the appearances of each couple of shots in training sports video. The following formulas embody the basic idea of derivation of Bi-gram.

$$\begin{cases} \mathrm{P}(h_j|h_i) = \alpha N(h_i, h_j)/N(h_i), & if \quad N(h_i) \neq 0 \\ \mathrm{P}(h_j|h_i) = 1/l, & otherwise \end{cases} \tag{11}$$

where $N(h_i, h_j)$ is the number of times shot h_j follows shot h_i and $N(h_i)$ is the number of times that shot h_i appears. l is the total number of distinct shot models, and α is chosen to ensure that $\sum_{j=1}^{l} \mathrm{P}(h_j|h_i) = 1$.

3.4 Procedure of Shot Segmentation and Classification

With Bi-gram and tri-shot HMM constructed, sports video can be segmented and classified into 3 types of shots. Since log operator can transform product operation into summation operation, in practice, product probability in the equations presented in Section 2 can be transformed into summation of corresponding log probability.

As mentioned in Section 2, a super HMM is obtained by concatenated the corresponding shot HMMs using a pronunciation lexicon. So each shot HMM in the super HMM is considered as a node. Hence, the task of shot segmentation and classification is to find a path of the maximum log probability from the start node to the end node in the super HMM, as shown in Fig.1.

For an unknown shot sequence of T SSUs whose feature vector sequence is $O(O = o_1 o_2 \ldots o_T)$, each path from start node to end node in the super HMM which passes through exactly T emitting HMM states is a potential recognition hypothesis. The log probability of each path is computed by summing the log probability of each individual transition in the path and the log probability of each emitting state generating the corresponding SSU. Intra-HMM transitions are determined from HMM parameters, and inter-HMM transitions are determined by Bi-gram [12].

Thus, the path of maximum log probability is the best result of shot segmentation and classification. As a result, the nodes (shot HMMs) on the best path are the optimal result of shot classification. Since each state in a shot HMM matches a SSU, the boundaries of each shot are the first and last SSUs of it, which realizes shot segmentation simultaneously.

4 Experimental Results and Analysis

We first evaluated the performance of our proposed statistical framework on soccer and badminton videos, and compared the performance of our method with that of a general two-stage method of shot segmentation and classification using the same features and test videos, and then we analyzed experimental results and parameters. Test video is CIF- $352 \times 288 \times 25$fps got from FIFA World Cup 2006 and 2005 World Badminton Championship including 4 full half-time soccer videos and 4 full game badminton videos. The implementation of the proposed framework is based on HTK 3.3 [12]. The ground-truth of boundaries and type of each shot are labeled manually. Table.1 shows the test corpus, and each shot lasts at least 1s, i.e. 25 frames.

Table 1. Test Videos

Name	Match(2006)	Length (min)	Name	Match(2005-8-21)	Length (min)
Soccer1	GER-ITA(7-4)	46:42	Badminton1	INA-THA(Mixed Doubles)	24:11
Soccer2	ENG-POR(7-1)	46:38	Badminton2	MAS-INA(Men's Singles)	26:12
Soccer3	GER-ARG(6-30)	46:05	Badminton3	CHN-INA(Mixed Doubles)	26:24
Soccer4	SUI-UKR(6-26)	47:04	Badminton4	NZL-CHN(Mixed Doubles)	17:09

Results are assessed by recall and precision rate which can be computed by

$$\text{Recall} = \text{Correct}/\text{Ground\_truth} \qquad \text{Precision} = \text{Correct}/\text{Detected} \tag{12}$$

where Detected is the shot number obtained by shot segmentation, and Correct is the number of correctly classified shots. Correct classification denotes not only the shot type is correctly recognized but also the overlap of the ranges of classified shot and actual shot more than 90 percent of the length of actual shot.

The proposed framework is tested on soccer and badminton videos since they are complete different types. Experimental results are shown in Table 2.

Table 2. Experimental results on soccer and badminton video

Shot	Ground-truth		Detected		Correct	
Type	soccer	badminton	soccer	badminton	soccer	badminton
LS	570	156	658	165	559	152
MS	392	103	474	121	326	80
CS	402	215	395	266	313	212
Total	Soccer: Recall = 87.8%; Precision = 78.5% Badminton: Recall =93.7% ; Precision = 80.4%					

On the average, the proposed framework achieves 87.8% recall and 78.5% precision rate on soccer videos, and 93.7% recall and 80.4% precision rate on badminton videos. The results are promising, which demonstrates the effectiveness of our general framework for shot segmentation and classification. We studied that false alarms are mainly caused by the misclassification of CS and MS and over segmentation of CS and MS, and the performance can be improved by applied more complicated features instead of simple color and motion features.

Experiments Using SVM. we applied a general method proposed in [5] for shot segmentation and classification using the same features for comparison. To simplify the procedure, we use SVM to classify manually segmented shots in above soccer videos. We chose $C = 2^1$ and $\gamma = 2^{-10}$ by cross-validation, and the experimental results are shown in Table 3. The total precision rate is 70.3%, which is far less than 78.5% precision rate achieved by our method. Note that shot classification using SVM is performed on the manual segmented shots, which avoids occurrences of wrong classifications caused by inaccurate segmentation. Thus, the proposed framework performs much better.

Table 3. Experimental results of SVM classification

Shot Type	Ground-truth	Correct	Precision(%)	Total
LS	570	530	92.9	Precision = 70.3%
MS	392	181	46.2	
CS	402	248	61.7	

Experiments Using Various Orders of Difference Features. Another 2 types of feature vectors are tested to demonstrate that difference can improve the performance, and they are 6-D feature vector (without difference feature), 12-D feature vector (with 1^{st} difference feature). As we can see from Table 4, performance is significantly improved with higher order difference features.

Table 4. Experimental results using various orders of difference features

Name	6-D feature vector	12-D feature vector	18-D feature vector
	Recall : Precision (%)	Recall : Precision (%)	Recall : Precision (%)
Soccer1	83.5 : 72.5	85.8 : 74.7	86.1 : 76.2
Soccer2	84.1 : 73.2	87.3 : 74.1	87.9 : 78.9
Soccer3	83.2 : 68.8	87.9 : 77.1	90.1 : 79
Soccer4	80.6 : 75.6	85.2 : 78.5	86.7 : 80

Experiments Using Various Combinations of SSU and SR. As described in Section 3.1, the size of SSU and SR determined the fineness and number of feature vector in a shot, respectively. 8 groups of different SSU and SR are tested on soccer video using 18-D feature to find the best combination of SSU and SR.

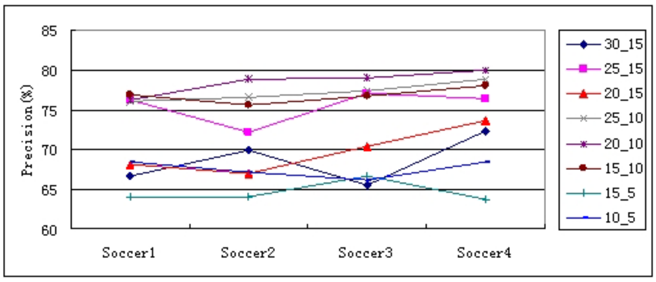

Fig. 4. Experimental results using various combinations of SSU and SR

Results are shown in Fig. 4, where the combination of SSU and SR is denoted by SSU_SR (frames). As we can see, performances are much better when SR is more than 10 frames and the ratio of SSU to SR is in [1.5, 2]. So in our work, the sizes of SSU and SR are set as 20 and 10 frames, respectively.

5 Conclusions

A statistical framework for shot segmentation and classification in sports video is presented in this paper. The main idea of the proposed framework is that the task of shot segmentation and classification is taken as a conditional probability which implicates intra-shot and inter-shot information. Thus, the proposed method realizes shot segmentation and classification simultaneously, and achieves much better performance than general two-stage method on soccer videos. Experimental results on badminton videos are also promising, which demonstrate the framework can be extended to other sports video. Furthermore, difference of feature introduced from speech recognition area is applied in our framework and has been proved it's superiority in improvement of performance.

In the further work, we will employ higher semantic features to enhance the performance, and apply the framework to event detection in sports video.

References

1. Hanjalic, A.: Shot-boundary Detection: Unraveled and resolved? IEEE Trans. Circuits and Systems for Video Technology 12, 90–105 (2002)
2. Lexing, X., Peng, X., Chang, S.-F., Divakaran, A., Sun, H.: Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models. Pattern Recognition Letters 24(15), 767–775 (2003)
3. Ekin, A., Tekalp, A.M., et al.: Automatic Soccer Video Analysis and Summarization. IEEE Trans. Image Processing 12, 796–807 (2003)
4. Dahyot, R., Rea, N., Kokaram, A.: Sports Video Shot Segmentation and Classification. In: SPIE Int. Conf. Visual Communication and Image Processing, pp. 404–413 (2003)
5. Wang, L., Liu, X., Lin, S., Xu, G., Shum, H.-Y., et al.: Generic Slow-motion Replay Detection in Sports Video. IEEE ICIP, 1585–1588 (2004)
6. Duan, L.-Y., Xu, M., Tian, Q.: Semantic Shot Classification in Sports Video. In: SPIE Proc. Storage and Retrieval for Media Databases, pp. 300–313 (2003)
7. Duan, L.Y., Xu, M., Tian, Q., et al.: A Unified Framework for Semantic Shot Classification in Sports Video. IEEE Trans. on Multimedia 7, 1066–1083 (2005)
8. Xu, M., Duan, L., Xu, C., Tian, Q.: A fusion scheme of visual and auditory modalities for event detection in sports video. IEEE ICASSP 3, 189–192 (2003)
9. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceeding of the IEEE 77, 257–286 (1989)
10. Ney, H., Ortmanns, S.: Progress in dynamic programming search for LVCSR. Proceeding of the IEEE 88, 1224–1240 (2000)
11. Bilmes, J.: A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report of University of Berkeley, ICSI-TR-97-021 (1998)
12. Young, S., Evermann, G., et al.: The HTK book (for HTK version 3.3). Cambridge University Tech Services Ltd. (2005), http://htk.eng.cam.ac.uk/