CLASSIFICATION OF TV PROGRAMS BASED ON AUDIO INFORMATION USING HIDDEN MARKOV MODEL

Zhu Liu, Jincheng Huang and Yao Wang
Department of Electrical Engineering
Polytechnic University, Brooklyn, NY 11201
{zhul, jhuang, yao}@vision.poly.edu

Abstract - This paper describes a technique for classifying TV broadcast video using Hidden Markov Model (HMM) [1]. Here we consider the problem of discriminating five types of TV programs, namely commercials, basketball games, football games, news reports, and weather forecasts. Eight frame-based audio features are used to characterize the low-level audio properties, and fourteen clip-based audio features are extracted based on these frame-based features to characterize the high-level audio properties. For each type of these five TV programs, we build an ergodic HMM using the clip-based features as observation vectors. The maximum likelihood method is then used for classifying testing data using the trained models.

INTRODUCTION

Video sequence is a rich multimodal information source, containing audio, speech, text, image and motion, etc. Efficient indexing and retrieval of video is becoming extremely important, which requires automatic understanding of semantic content. Audio as a counterpart of visual information in video sequence got more attention recently for its semantic content discrimination capability [2-4]. Improvement in the segmentation and classification of video sequence was reported by using both visual and audio information. HMM has good capability to grasp the temporal statistical property of stochastic process and is used widely in pattern recognition field. Recently Boreczky [5] used HMM framework for video segmentation using audio and image features. The emphasis of this paper is on applying the HMM for video content classification using audio information. The video sequence is first manually segmented such that each sequence only contains one kind of TV programs. Then clip-based audio features are extracted and used to train the HMMs of the five TV programs. The classification results and related analysis are reported in this paper.

AUDIO FEATURES DESCRIPTION

The audio signal is sampled at 22050 Hz and 16 bits/sample. The audio stream is segmented into clips that are 1.5 seconds long with 1 second overlapping with the previous ones. Each clip is then divided into frames that are 512 samples long and are shifted by 256 samples from the previous frames. For each frame, we extract eight short time features: 1) root mean square (RMS) volume; 2) zero crossing rate (ZCR); 3) pitch period (the search range is 2.3 ~ 15.9 ms); 4) frequency centroid; 5) frequency bandwidth; 6~8) energy ratio in three different subbands. The ranges of these three subbands are 0~630 Hz, 630~1720 Hz and 1720 ~ 4400 Hz. Based on these features we compute 14 clip-based features:

1) Non-silence ratio (NSR): the ratio of silent frames (decided by preset threshold) to the entire clip.
2) Volume standard deviation (VSTD).
3) Standard deviation of zero crossing rate (ZSTD).
4) Volume dynamic range (VDR): the difference of maximum and minimum volume of a clip normalized by the maximum volume in that clip.
5) Volume undulation (VU): the accumulation of the difference of neighbored peaks and valleys of the volume contour.
6) 4 Hz modulation energy (4ME): the frequency component around 4Hz of the volume contour.
7) Standard deviation of pitch period (PSTD).
8) Smooth pitch ratio (SPR): the ratio of frames that have similar pitch period as the previous frames (the difference is less than 0.68 ms) to the entire clip.
9) Non-pitch ratio (NPR): the ratio of the frames that no pitch is detected in the search range to the entire clip.
10) Frequency centroid (FC): energy weighted mean of frequency centroid of each frame.
11) Frequency bandwidth (BW): energy weighted mean of frequency bandwidth of each frame.

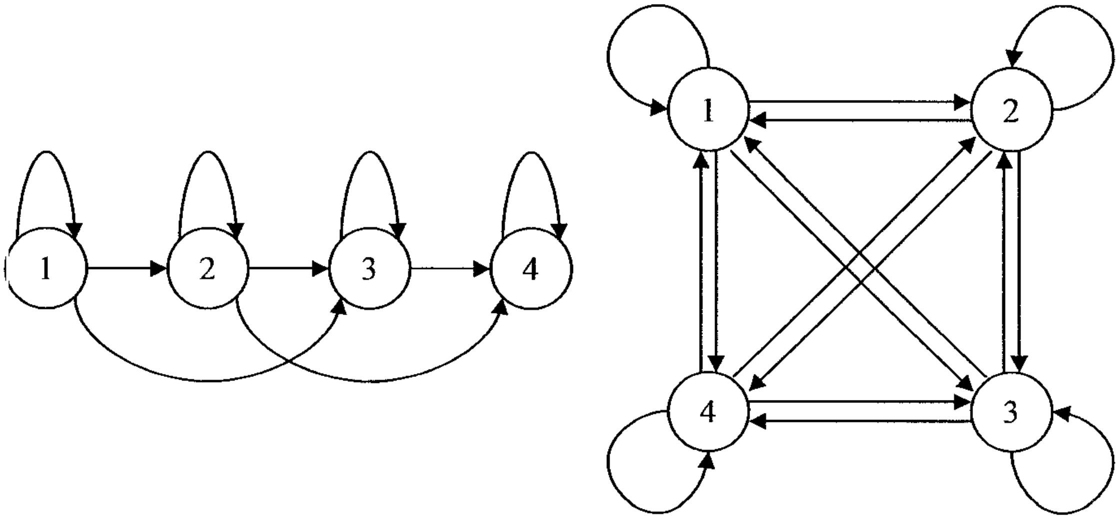
12-14) Energy ratio of subband 1-3 (ERSB1-3): energy weighted mean of energy ratio in subband 1-3 of each frame.

Since the dynamic ranges of these features differ a lot, we normalize them by their standard deviations that are computed based on the training data. For more detailed description on audio features, see [4].

HMM CLASSIFICATION

A discrete HMM is determined by three groups of parameters: the state transition probability $\mathbf{A}=\{a_{ij}\}$, where $a_{ij}=P(q_{t+1}=j|q_t=i)$; the observation symbol probability $\mathbf{B}=\{b_j(k)\}$, where $b_j(k)=P(\mathbf{o}_t=\mathbf{v}_k|q_t=j)$; and the initial state distribution $\pi=\{\pi_i\}$, where $\pi_i=P(q_1=i)$. Here, q_t is the state at time t, $\mathbf{v}_k$ is the distinct observation symbols in observation space and $\mathbf{o}_t$ is the observation vector in time t. For convenience we use $\lambda=(\mathbf{A}, \mathbf{B}, \pi)$ to indicate the model parameters. In our case, the observation space is the feature space and we need to quantize it to a finite number of vectors before we utilize the discrete HMM. Here we generate the codebook using binary split algorithm described in [1].

HMM has been successfully applied in several large-scale laboratory and commercial speech recognition systems. In traditional speech recognition system, a distinct HMM is trained for each word or phoneme, and the observation vector is computed every frame (10 ~ 30 ms). Here we do not need to grasp the detail information at the resolution of several milliseconds. We are interested in the semantic content that can only be determined over a longer time duration. Based on this consideration, we compute the feature vector for every clip. Another major difference is the state transition probability **A**. In speech recognition, a phoneme or a word has a well defined temporal structure so that some states cannot be reached from some other states and the corresponding a_{ij} are zero. Figure 1 (a) shows a 4-state left-right model that is suitable for speech recognition. In our case, the temporal structure can be repeated in the video sequence. For example, in the news report, a live report usually comes after anchorperson's introduction and then returns to the studio report. This kind of pattern can be repeated for several times in a single news program. Another obvious example occurs in games such as the football game, where there are a lot of cycles of attacks and pauses. Such temporal structures of video sequence require us to use ergodic HMM shown in figure 1 (b), where each state can be reached from other sates and can be revisited after leaving.



(a) 4-state left-right HMM

(b) 4-state ergodic HMM

Figure 1. Illustration of two typical HMMs.

The HMM training process follows the Baum-Welch method [1]. The initial parameters of **A** and **B** are chosen randomly and the initial values of π are uniformly distributed for each state. After training we get λ_1, λ_2, ... λ_C, where C is the number of video classes. We use maximum likelihood method to classify the TV programs. For each testing sequence **o**, we compute the probability $P(\mathbf{o} \mid \lambda_i)$, i = 1 ,..., C. Then we can classify the testing sequence to the class with maximum probability. We can also use Viterbi algorithm to get the optimal state sequence underlying the testing sequence. This information is useful in determining the physical meaning of each state.

For the ergodic HMM, the initial state distribution of one state is also the probability of occurrence of that state in the model. **A** matrix shows us whether there exist frequent transitions from one state to others or not. Normally, each state has higher probability to stay at the same state than to transit to other states. If there exists a group of states that have low transition probabilities to stay at the same states and high probabilities to transit to each other, the number of states may be higher than necessary for HMM to model the training data. This group of states can merge to one state without influencing the performance. Through analyzing the **B** matrix, we can also find whether the number of states is enough for certain class. If the observation symbol probabilities of a set of states are similar, these states can also be merged to one state. If the observation symbol probability of one state is almost uniformly distributed, it means such state is not very useful and we can reduce the number of states for such class.

SIMULATION RESULTS

We collected video sequences from TV programs containing the following five scene classes: news reports, weather forecasts, commercials, live basketball games, and live football games. For each scene class, we collected 20 minutes video from different TV channels. These data are divided into two sets: training and testing data sets. The training data set includes 10 minutes audio for each scene class and the remaining 10 minutes audio in each class forms the testing data set. For each digitized audio sequence, we first compute clip-based feature vectors and then we use every 20 continuous clips as training sequence, with a shift of 1 clip between each two sequences. Such shifting of training data to generate training sequences is feasible because we are using an ergodic HMM. In such way we can improve the efficiency of limited training data. Totally we use 1000 sequences for training the model and 1000 sequences for testing for each kind of TV programs.

Since there is no simple theoretically correct way to choose the number of states and the number of observation symbols, we tried several cases of different combinations of them. Table 1 gives the overall classification accuracy for the different choices of state and symbol numbers. Here the overall classification accuracy is defined as the average classification accuracy for the five audio classes. From the table, we know the HMM classifier give best performance for our task when the number of states is 5 and the number of symbols is 128. The corresponding overall classification accuracy is 84.7%.

Number of symbols	Number of states				
	4	5	6	7	8
8	61.4	66.9	67.9	69.5	67.5
16	70.5	68.4	70.4	73.2	70.4
32	75.4	78.5	77.4	78.8	77.2
64	80.3	76.9	80.3	76.3	79.3
128	84.2	84.7	84.6	84.5	83.9
256	84.2	84.6	84.5	84.3	82.0

Table 1. Classification results (unit: 100%) for different state and symbol number.

Tables 2 and 3 give the state transition probability matrices of λ_1 (commercial HMM) with 5 and 6 states respectively. For 5-state λ_1, the diagonal items are much higher than the rest. On the contrary, for 6-state λ_1, the state 1 and state 5 have very low probability to stay at their own states but high probability to transit to each other. We also find that the observation symbol probabilities of these two states are quite similar. This explains that we get good result for 5-state HMM.

The worse performance for higher state may be due to the limited size of training data set.

State	1	2	3	4	5
1	0.9285	0.0074	0.0151	0.0246	0.0244
2	0.0038	0.9316	0.0187	0.0129	0.0330
3	0.0199	0.0214	0.9092	0.0407	0.0088
4	0.0535	0.0000	0.0000	0.8551	0.0914
5	0.0538	0.0227	0.0762	0.0480	0.7993

Table 2. State transition probability of 5-state HMM for commercial.

State	1	2	3	4	5	6
1	0.0000	0.0580	0.0000	0.0000	0.8602	0.0818
2	0.0469	0.8938	0.0000	0.0000	0.0320	0.0273
3	0.0000	0.0052	0.8896	0.0957	0.0000	0.0095
4	0.0001	0.0000	0.0649	0.8578	0.0447	0.0325
5	0.8868	0.0150	0.0275	0.0706	0.0000	0.0001
6	0.0755	0.0060	0.0150	0.0274	0.0227	0.8534

Table 3. State transition probability of 6-state HMM for commercial.

Table 4 gives the classification results for the HMM with 5 states and 128 symbols. From this table, we can see that the classifier can accurately distinguish among commercials, basketball/football games, and news/weather reports. But the separation of the news from weather reports is less successful. This is not surprising because they contain primarily pure speech. High level correlation information between successive clips that reflects the flow of the conversation may be necessary. The classification accuracy of commercials, games, news/weather reports is 93.4%. Using the neural network approach [4] on the same data sets, the overall classification accuracy is 72.8% and the classification accuracy of commercial, games, news/weather report is 86.8%. We get 11.9% improvement of overall classification accuracy by using HMM.

Data	Input Class				
Result	Commercial	Basketball	Football	News	Weather
Commercial	87.4	3.4	1.4	0.3	0.0
Basketball	2.5	85.4	8.4	1.2	0.0
Football	1.6	9.1	86.6	2.5	0.0
News	8.5	2.1	3.6	68.8	4.7
Weather	0.0	0.0	0.0	27.2	95.3

Table 4. Classification results (unit: 100%) of 5-state HMM with 28 symbols.

For the results presented here, the video sequences are manually segmented. We can also use the HMM to automatically segment the video data. For the purpose of segmentation, we can normalize the $P(\mathbf{o}\,|\lambda_i)$ with the length of observation sequence so that it will not always decrease with the length. By tracking the contour of P($\mathbf{o}$ |λ_i), we can find the points with maximum variation correspond the scene change points and the smooth parts of the contour corresponds to the same class.

This HMM framework can also be extended to utilizing the visual information. For example, we can extract features such as dominant colors and motion vectors for each frame and then extract clip-level features based on them. We can attach these visual features to existing audio features to create new codebook and train the HMM. We believe the visual information can further improve the classification accuracy within different games and news/weather reports. This is reasonable, for example, the background of football game is primarily green while in basketball games it is yellow or brown.

CONCLUDING REMARKS

In this paper, we have described a video content classifier based on HMM using audio features. Using ergodic HMM with 5 states and 128 symbols, we achieve 84.7% overall accuracy in classifying commercial, basketball game, football game, news, and weather forecast. The classification results reported here are meant to show the promise of applying HMM in video scene classification. Better results should be obtainable after combining with visual information and setting a prior known constraints to the models' parameters.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation through its STIMULATE program under Grant IRI9619114.

References

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

[2] J. Nam and A. H. Tewfik, "Combined Audio and Visual Streams Analysis for Video Sequence Segmentation," *Proc. of ICASSP'97*, Vol. 3, pp. 2665-2668, 1997

[3] C. Saraceno and R. Leonardi, "Audio as a Support to Scene Change Detection and Characterization of Video Sequences," *Proc. of ICASSP'97*, Vol. 4, pp. 2597-2600,1997.

[4] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," to appear in *Journal of VLSI Signal Processing System*, June 1998.

[5] J. S. Boreczky and L. D. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features", *Proc. of ICASSP'98*, Vol. 6, pp. 3741-3744, 1998.