JOINT VIDEO SCENE SEGMENTATION AND CLASSIFICATION BASED ON HIDDEN MARKOV MODEL

Jincheng Huang, Zhu Liu, and Yao Wang

Department of Electrical Engineering
Polytechnic University
Brooklyn, NY 11201
{jhuang, zhul, yao}@vision.poly.edu

ABSTRACT

Video classification and segmentation are fundamental steps for efficient accessing, retrieving and browsing large amount of video data. We have developed a scene classification scheme using a Hidden Markov Model (HMM)-based classifier. By utilizing the temporal behaviors of different scene classes, HMM classifier can effectively classify video segments into one of the predefined scene classes. In this paper, we describe two approaches for joint video classification and segmentation based on HMM, which works by searching for the most likely class transition path utilizing the dynamic programming technique.

1. INTRODUCTION

Video classification and segmentation are fundamental steps for efficient accessing, retrieving and browsing large amount of video data. Recently, several research groups have developed algorithms to detect scene change by incorporating audio and visual information. Most of these works [1, 2, 3] are based on some prior scene models, (e.g. dialog, setting, etc.) and accomplish scene segmentation and classification simultaneously. Such techniques are however very dependent on the particular definition of scene classes and are not easy to generalize.

Previously we have developed a scene segmentation scheme which identify scene breaks based on co-occurrence of significant changes in audio and visual characteristics[4]. We also developed a scene classification scheme using a Hidden Markov Model (HMM)-based classifier[5, 6]. By utilizing the temporal behaviors of different scene classes, HMM classifier can effectively classify video segments into one of the predefined scene classes, namely, commercial, basketball games, football games, news, and weather forecast. In the classification work, it is assumed that segmentation has been accomplished previously and the input segment to the HMM classifier belongs to only one scene class. The difference between our approach and some other works [2, 3] is that our segmentation scheme is based on general audio-visual features that are not tuned to identify particular scenes. Our classification scheme is on the other hand totally driven by probabilistic models generated based on training data instead of heuristic rules. The same approach can be applied to identify any scene type.

Our previous segmentation scheme is based on audio-visual features observed over a short interval, and the scene break is detected when the relative change of these features at any time instance exceeds a certain preset threshold. One problem with this approach is that it can lead to spurious breaks when a scene contains different segments that have very different audio-visual characteristics. Another difficulty lies in the selection of proper thresholds.

Here, we propose two approaches for joint video scene classification and segmentation by taking advantage of HMM. In the first approach, we compute the likelihood that a short video segment belongs to a particular scene class for every video segment. The likelihood values are used as intermediate values. We reconstruct an optimum scene transition path for the entire input video sequence that has the highest accumulative likelihood. In the second approach, we build a super HMM by concatenating HMM's for different scene classes. The scene class transition path is obtained by searching the path within a class and between classes using the dynamic programming technique.

The outline of the paper is as follows: In Sections 2 and 3, we describe the two approaches for joint video classification and segmentation based on the Hidden Markov Model. Simulation results are given in Section 4. Section 5 concludes the paper.

2. APPROACH I

In an ideal case in which each video short segment is correctly identified using an HMM classifier, the scene

This work was supported by the National Science Foundation (NSF) through its STIMULATE Program under Grant No. IRI-9619114.

transition can be easily spotted. However, in reality, the likelihood value for the correct class can be temporally plunged due to the mismatch between the audio-visual features over a short time interval and the model while the likelihood for some incorrect classes take over. An example is shown in Figure 1. At clips 420–450, the likelihood for the commercial model are lower than that for news and basketball models. However, this portion of program belongs to a commercial. One way to solve this problem is by basing the decision on an accumulative likelihood that a segment belongs to a particular scene class from the staring point, with a penalty assigned to a transition from one class to another to suppress false transitions. This idea can be realized by dynamic programming.

Here, we introduce four variables:

- $\mathcal{L}_t(j)$: accumulative likelihood for class j ends at time t;
- $\mathcal{A}_t(j)$: class backtrack pointer for class j ends at time t;
- $C(i,j)$: penalty for transition from class i to class j;
- $l_t(j)$: the likelihood value for class j at time t; This value is computed from the audio-visual feature vector computed over a fixed-length segment, known as an observation sequence.

The dynamic programming technique tries to find the optimum path, which maximizes the accumulative likelihood at every time t, such as, $\mathcal{L}_t(j) = \max_i\{\mathcal{L}_{t-1}(i) + C(i,j)\} + l_t(j)$. The search algorithm is outlined as following:

- Initialization:
$$\mathcal{L}_1(i) = l_1(i) \qquad \text{and} \qquad \mathcal{A}_1(i) = 0.$$
- Recursion:
$$\begin{aligned} \mathcal{L}_t(j) &= \max_{1\leq i\leq N}\{\mathcal{L}_{t-1}(i) + C(i,j)\} + l_t(j); \\ \mathcal{A}_t(j) &= \arg\max_{1\leq i\leq N}\{\mathcal{L}_{t-1}(j) + C(i,j)\}. \end{aligned}$$
- Termination:
$$\mathcal{L}^* = \max_{1\leq i\leq N}\mathcal{L}_T(i) \quad \text{and} \quad C_T^* = \arg\max_{1\leq i\leq N}\mathcal{L}_T(i)$$
- Class Backtracking:
$$C_t^* = \mathcal{A}_{t+1}(C_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1.$$

3. APPROACH II

In this approach, we use a technique developed for speech recognition. To hypothesize a scene class sequence $C = c_1, \cdots, c_L$, where L is the number of segments, we can imagine a super HMM that is obtained by concatenating the HMM's for different classes. The search space can be described as a huge network where the best state transition path has to be found. The search has to be performed at two levels: at the state level and at the class level. The paths at the two levels can be searched efficiently by dynamic programming.

We use the one-pass dynamic programming search [7] to find the optimal class sequence for a given observation sequence $X = \{x_1, x_2, \cdots, x_T\}$. The algorithm requires two arrays:

- $Q(t,s;c)$: score of the best path up to time t that ends in state s of class c.
- $B(t,s;c)$: start time of the best path up to time t that ends in state s of class c.

As illustrated in Figure 2, the path is searched within the class and among the classes. Within the class, the recurrence equation is as following:

$$\begin{aligned} Q(t,s;c) &= \max_{0\leq s'\leq N(c)}\{p(x_t, s|s';c)\cdot Q(t-1,s';c)\}, \\ B(t,s;c) &= B(t-1, s_{max}(t,s;c);c), \qquad 1\leq s\leq N(c), \end{aligned}$$

where $N(c)$ is number of states in class c, $s_{max}(t,s;c)$ is the optimum predecessor state for the hypothesis $(t,s;c)$, *i.e.*, $s_{max}(t,s;c) = \arg\max_{s'}\{p(x_t,s|s';c)\cdot Q(t-1,s';c)\}$.

To hypothesize the potential scene class boundary, the termination quantity $(H(c;t))$, a class traceback pointer $(R(c;t))$, and a time traceback pointer $(F(c;t))$ are introduced as:

$$\begin{aligned} H(c;t) &= \max_{1\leq b\leq K, b\neq c}\{p(c|b)\cdot Q(t,S_b;b)\}, \\ R(c;t) &= \arg\max_{1\leq b\leq K, b\neq c}\{p(c|b)\cdot Q(t,S_b;b)\}, \\ F(c;t) &= B(t, S_b, R(c;t)), \\ \text{with} \qquad S_b &= \arg\max_{1\leq s\leq N(b)} Q(t,s;b), \end{aligned}$$

where K is the total number of reference classes, $p(c|b)$ is the class transition probability of class b to class c. To allow for successor classes to be started, a special state $s=0$ is introduced and passed on both the score and the time index:

$$\begin{aligned} Q(t-1, s=0;c) &= H(c;t-1) \\ B(t-1, s=0;c) &= t-1. \end{aligned}$$

Figure 2 (a) illustrates the time alignment, which gives the optimal scene sequence. In this example, the sequence is $(2, K, 1)$ and transitions occur at t_1 and t_2. As illustrated in Figure 2 (b) , $Q(t,s;c)$ and $B(t,s;c)$ are determined for every state within each class at every time instance t. Then, $H(t;c)$ is computed, and $R(t;c)$ and $F(t;c)$ are recorded for all K classes. Before compute $Q(t,s;c)$, the score value and the backtrack time value for the potential scene change $Q(t-1, s=0;c)$ and $B(t-1, s=0;c))$ are set. The solid dot in Figure 2 (b) indicates a possible scene transition for state 2 at class

b to class c. The process starts at time $t = 1$ and ends at $t = T$ in a strictly left-right fashion. When time T is reached, the optimum scene sequence C_l^* and the time for the scene transition T_l^* can be found by tracing back $R(c;t)$ and $F(c;t)$, respectively, as following

$$\begin{aligned}
C_L^* &= \arg\max_c Q(T, S_c; c) \\
&\quad \text{with} \quad S_c = \arg\max_s Q(T, s; c); \\
T_L^* &= F(T, C_L^*); \\
C_l^* &= R(T_{l+1}^*, C_{l+1}^*), \qquad l = L-1, \ldots, 1; \\
T_l^* &= F(T_{l+1}^*, C_{l+1}^*), \qquad l = L-1, \ldots, 1.
\end{aligned}$$

4. SIMULATION RESULTS

As described in our early studies [5, 6], we focus on five scene classes: commercial, live basketball game, live football game, news, and weather forecast. We gathered 10 minutes of video for each scene class to train the HMM parameters. We also digitized several video segments which included various scene transitions, such as from news to commercials, basketball game to commercials. In the preliminary study, we used the fourteen audio features developed for scene classification [5, 6]. The features are extracted for every audio clip, which has length of 1.5 seconds and is overlapped with the previous clip with 1 second. 5-state Ergodic HMM's with 256 observation symbols are used for each scene class.

The testing sequence we present here includes portions of news program and commercials. It begins with a news program and then change to commercial at clip 271. Figure 1 shows the result using Approach I. The likelihoods are calculated for every observation sequences with length of 20 clips. The next sequence is one-clip shifted from the current one. The transition from news to commercials was detected at clip 304 while the true transition is at clip 271. The misclassification is due to the dominant speech signal present in the portion from clips 271 to 304. Notice that at clips 420–450 (in the commercial potion), the likelihood values for the basketball and the news models are higher than those from the commercial model. By using dynamic programming, this portion is correctly identified as commercial.

With Approach II, the transition was detected at clip 317, which is similar to the result from Approach I. Because the data for collecting scene transition probabilities are limited, these probabilities are obtained by trial and error in the current simulation. Larger transition probabilities cause fluctuation among classes while smaller one tends to lead to a delay in detecting transition.

5. CONCLUSION

We proposed two approaches for joint video scene segmentation and classification based on Hidden Markov Model. Both approaches gave the similar result. The implementation of Approach I is simpler but it requires more computation because each observation sequence is overlapped with the previous one. Indeed, this approach can improve the scene classification accuracy, in which the noisy classification for each short sequence can be eliminated (as at clips 420-450 in the testing sequence). The Approach II can be more accurate because it makes decision at the clip level, but for the same reason, it can lead to a noisier segmentation.

6. REFERENCES

[1] A. G. Hauptman, M. J. Witbrock, "Story Segmentation and Detection of Commercials in Broadcast News Video," in *Advances in Digital Libraries Conference (ADL-98)*, (Santa Barbar, CA), April 22–24, 1998.

[2] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene Determination based on Video and Audio Features," Technical Report TR-98-020, Praktische Informatik IV, University of Mannhein, Nov. 1998.

[3] C. Saraceno and R. Leonardi, "Audio as a Support to Scene Change Detection and Characterization of Video Squence," in *Proc. of International Conference on Acoustic, Speech, and Signal Processing*, pp. 2597-2600, 1997.

[4] J. Huang, Z. Liu, and Y. Wang, "Integration of Audio and Visual Information for Content-based Video Segmentation," in *IEEE International Conference on Image Processing (ICIP98)*, vol. 3, (Chicago, IL), pp. 526–530, Oct. 1998.

[5] Z. Liu, J. Huang, and Y. Wang, "Classification of TV Programs Based on Audio Information using Hidden Markov Model," *IEEE Workshop on Multimedia Signal Processing*, pp. 27–32, Log Angeles, CA, Dec. 7–9, 1998.

[6] J. Huang, Z. Liu, Y. Wang, Y. Chen, E. K. Wong, "Integration of Multimodal Features for Video Classification based on HMM," *IEEE 1999 Workshop on Multimedia Signal Processing (MMSP99)*, pp. 53–58, Copenhagen, Denmark, Sept. 13–15, 1999.

[7] H. Ney and S. Ortmanns, "Dynamic Programming Search for Continuous Speech Recognition," *IEEE Signal Processing Magazine*, pp. 64–83, Sept 1999.

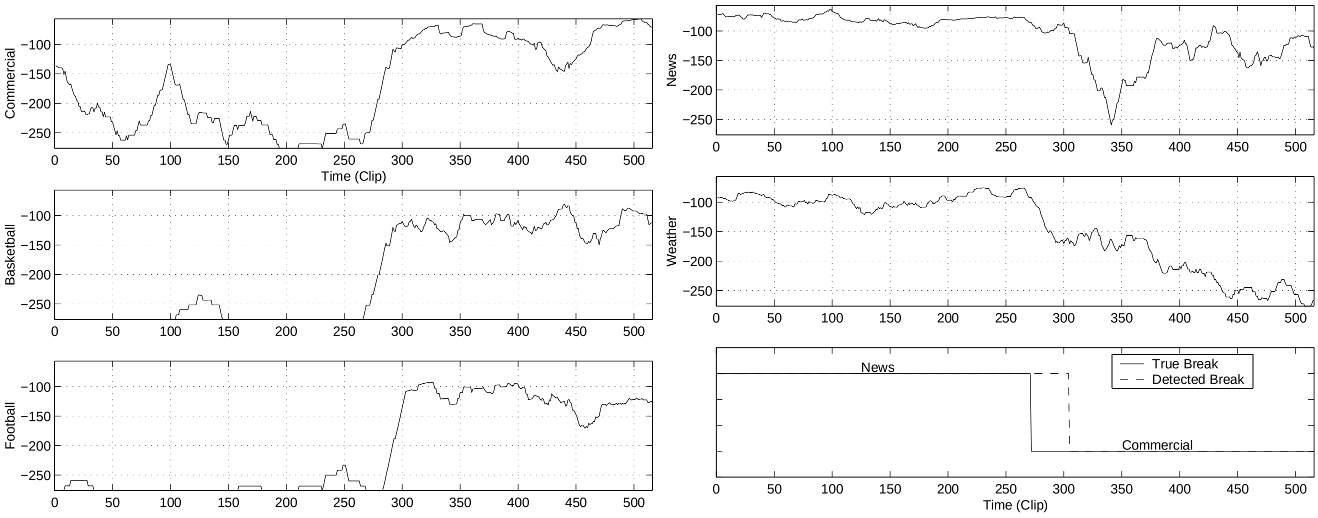
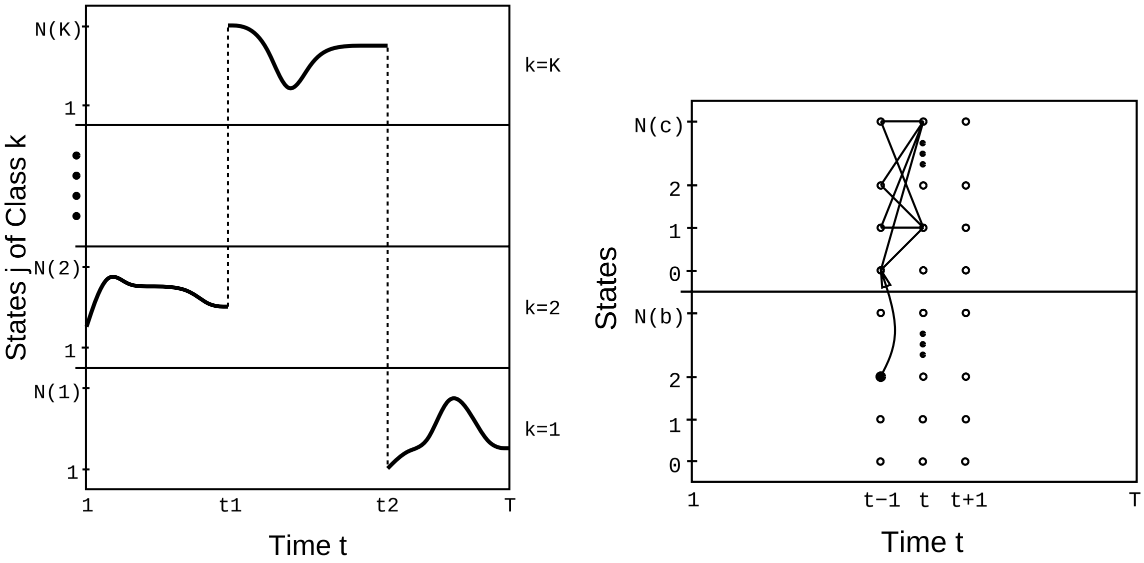


Figure 1: Segmentation and classification result using Approach I. The lower-right graph shows the detected scene break *vs* the true break. Other five graphs are the likelihoods resulting by individual HMM classifiers.



(a)

(b)

Figure 2: (a) Illustration of path alignment of an optimal scene sequence; (b) Illustration of path combinations (within class and between classes) for one-pass dynamic programming search