COMPARATIVE ANALYSIS OF HIDDEN MARKOV MODELS FOR MULTI-MODAL DIALOGUE SCENE INDEXING

A. Aydın Alatan, Ali N. Akansu and Wayne Wolf[†]

New Jersey Center for Multimedia Research
New Jersey Institute of Technology & [†]Princeton University
e-mail: alatan@oak.njit.edu

ABSTRACT

A class of audio-visual content is segmented into dialogue scenes using the state transitions of a novel hidden Markov model (HMM). Each shot is classified using both audio track and visual content to determine the state/scene transitions of the model. After simulations with circular and left-to-right HMM topologies, it is observed that both are performing very good with multi-modal inputs. Moreover, for circular topology, the comparisons between different training and observation sets show that audio and face information together gives the most consistent results among different observation sets.

1. INTRODUCTION

Digital multimedia content has been grown enormously within the last decade. Management of this huge content with limited computational power and networking capacity requires sophisticated algorithms. These algorithms are required to extract the descriptions of data for different applications.

A typical application is of finding conversations of people in multimedia content. Among major multimedia content classes [1], fiction entertainment class includes mostly post-edited content, such as movies, TV-series and sitcoms. In this class, the stories are presented to the viewers after they are edited by the experts in the content production area. This paper presents a novel method which determines the dialogue scenes from a content in fiction entertainment class. Although, every multimedia data might contain human communications (i.e. dialogues) in its content, the proposed model is more appropriate for motion pictures or TV series, which are basically, post-edited after recording for improving the presentation of the events (dialogues, actions) to the viewers.

Dialogue scene analysis is especially necessary for abstraction purposes. Rather than assigning key-frames for each shot of a video clip, it is more meaningful and efficient to represent the clip by a number of key-frames or short trailers for each dialogue scene. Moreover, the quantitative analysis between the durations of dialogue and non-dialogue scenes might also give an idea about the genre of the movie, as well.

2. SCENE ANALYSIS

While a *shot* is usually defined as a single continuous recording of a camera, a *scene* consists of concatenation of shots (combined during editing stage), possibly recorded at the same location and presenting a meaningful part of the whole story. Finding semantically meaningful parts of the story (i.e. scene analysis) is a necessary (also a challenging) step in multimedia indexing. Such an analysis is an important step from low-level features, such as color, motion, voice pitch, to the high-level semantic descriptions of the content.

2.1. Multi-modal Scene Analysis

Multi-modal analysis combines different components belonging to the same information content. In other words, for a movie, one might integrate video, audio, closed-caption or textual overlay information together for different applications. Although, it has obvious advantages over uni-modal analysis, the fusion of data with different nature is still a difficult problem to solve.

Multi-modal approaches have been investigated for shot-boundary detection [2], speaker-dependent temporal indexing [3], story unit identification [4] and violent scene detection [5]. The simulation results of all these techniques point out an improvement over uni-modal counterparts, as expected.

2.2. Dialogue Scene Analysis

Although it is very difficult to give a precise definition, a *dialogue scene* can be defined as a set of consecutive shots which contain conversations between a number of people. There are two challenging goals for the dialogue scene analysis : determining a dialogue as a whole (neither totally missing it nor erroneously dividing it into multiple sub-dialogues) and finding the boundaries of a dialogue (which can be quite subjective in some cases) with a good accuracy.

It is also possible that some shots in a dialogue scene do not contain a conversation or even a person, but it should still be included in the dialogue due to semantic coherence. For example, while two people are talking to each other, the shots usually contain the faces and speech content of these people. In order to emphasize another subject, the director of the movie may insert shot(s) of another object or location between these shots. Since the conversation continues after this short interrupt, although visually irrelevant, but semantically relevant shot(s) should also be included into the dialogue scene, as well. However, such random events make the analysis more difficult due to their indeterministic nature.

The analysis of a dialogue scene is usually achieved by using either a deterministic or a probabilistic approach. While the deterministic methods usually cluster consecutive shots by utilizing appropriate measures [1, 4], the probabilistic methods use hidden Markov models (HMM) to represent their states (i.e. scenes) in the content [6]. Although, both approaches are still quite immature and there is no fair comparison yet, HMM-based formulation looks more promising due to two factors [6]. The first reason lies within the random behavior of any natural language which does not permit to put a number of deterministic rules to model this behavior. As the second reason, following the analogy between spoken and visual languages (cinema), statistical models are also expected to parse the shots of a movie (like the phonemes of a word) better compared to non-statistical approaches. Moreover, clustering is not flexible enough to parse video programs into useful syntactic structures [6].

3. HMM-BASED MULTI-MODAL DIALOGUE SCENE ANALYSIS

Considering the performance of multi-modal video indexing and HMM-based dialogue modeling, we propose a novel method that detects the dialogue scenes in a movie using a multi-modal HMM-based approach.

3.1. The Elements of a Dialogue Scene

A dialogue scene requires three elements at the same time; some *people*, their *conversation* and a *location*. The importance of these three components might differ. Intuitively, the conversation between characters is the most important among the three elements. On the other hand, a common location might not be necessary for a phone conversation scene, hence it is the least defining factor.

The elements of a dialogue must be extracted from the multimedia content. The simplest way to detect a conversation is to use the audio track of an audio-visual content. Since it is common to show the face of speakers during a dialogue, face analysis is a good choice to determine the presence of people within shots. Finally, the location information usually requires some simple visual clues.

3.2. Dialogue Scene Analysis Hierarchy

Multi-modal dialogue scene analysis can be categorized into a three-level hierarchy according to the (computational) complexity of the resulting systems. This classification is shown in Table 1.

Level	Audio Analysis	Visual Analysis	Resulting system
Low	`silence/ no-silence`	`shot info`	`shot detection`
Mid	`silence/ music/ speech`	`shot info + face/noface + old/new locat.`	`dialogue scene detection`
High	`silence/ music/ speaker1...N`	`shot info + face1...M locat1...K`	`dialogue scene description`

Table 1: Classification of multi-modal scene analysis according to computational complexity levels of the algorithms.

The applications, which require considerably lower complexity, may simply use methods which have silence/no-silence segmentation of the audio track data and a set of visual shot boundaries. An improvement is shown over shot-boundary detection compared to using only visual data [2]. A method from high-level analysis class possibly requires robust unsupervised speaker, face and location classification algorithms, in order to segment speech signals into N speakers in a soundtrack, to classify faces of M different people in a movie, and to detect K locations in which the events take place, respectively. Considering the state-of-the-art audio-visual analysis methods, all these three goals are computationally quite complex. Moreover, they might not be robust against generic scenes and soundtracks.

In this paper, we focus on medium-complexity techniques. These methods require only a classification between silence, speech and music data for their audio analysis part. This problem is relatively easier to solve [4]. For the visual part, we only need to find whether a face exists at each shot or not. Face detection is also a mature topic with robust solutions [6, 3]. Finally, location analysis can be simplified into detection of significant changes between shots using simple histogram-based methods [7, 2, 5].

3.3. Hidden Markov Modeling of Dialogue Scenes

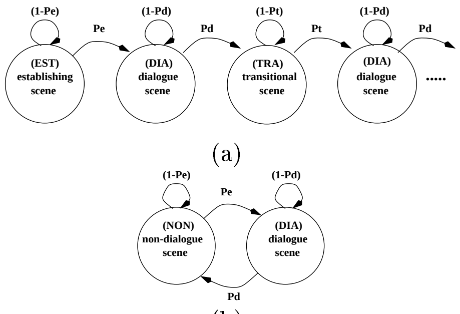
Hidden Markov Models (HMM) are powerful statistical tools which have been successfully utilized in speech recognition and speaker identification fields [8]. Recently, they have also found applications in content-based video indexing area for solving video scene segmentation [6] and shot-boundary detection [7] problems.

HMM-based analysis usually requires answers to the questions, below. While the first three questions depend on a particular design, the last two problems have general solutions in the literature [8].

- What are the hidden states?
- How are these states connected (i.e. topology) ?
- How do we observe the hidden states?
- How do we obtain the statistical parameters of HMM?
- How do we find a state sequence for an observation data set?

Assigning the scenes of the content to the states of the HMM is the most straightforward approach. According to the scene classifications of content producers, the states of the model can also be determined. Dialogue, action, establishing and transition are the mostly employed scenes by producers.

The HMM states can be related with each other using different HMM topologies. One possibility is a left-to-right HMM state diagram (Fig. 1.a) while another is a circular HMM (Fig. 1.b). Left-to-right state diagram simply tells that movies start with a state, called establishing scene, which is followed by a dialogue scene. Once a dialogue scene ends, we enter into a transitional (or establishing) scene in which a conversation does not exist and this order repeats itself. As an obvious drawback, left-to-right model strictly requires knowing the number of scenes (states) in the content, beforehand. On the other hand, without such a constraint, circular topology can be constructed from only two states, representing dialogue and non-dialogue scenes.



(a)

(b)

Figure 1: (a) Left-to-right and (b) circular HMM state diagram for modeling dialogue scenes in movies.

At each HMM state, there are some observable output symbols with some associated probabilities. The observable symbols are important factors which determine the performance of the model. For HMM-based dialogue scene analysis, the output symbols can be obtained by merging combinations of conversation, face and location descriptions.

Once the model topology and observation symbol set are determined, the next step is of training the model by some data and determining the initial, state-transition and observation probabilities. Baum-Welch algorithm [8] is the well-known solution for training HMMs. After the model probabilities are obtained by using the training data, the observations of a test data are fed into a dynamic programming routine in order to obtain the best state-sequence, i.e. the sequence of scenes in the video clip.

4. SIMULATIONS

Simulations are conducted to compare different topologies, different observation symbol sets and the effects of different training data on HMM-based dialogue scene analysis. Since the extraction process is not a factor for these comparisons, it is assumed that all such low level descriptors are available. In other words, audio and video features are obtained manually from training and test data sets.

4.1. Simulated system

The block-diagram of the simulated system is shown in Fig. 2. As a first step, it is necessary to decode and demultiplex the compressed input bit-stream (consisting of audio and video data) into its audio and video components. While video signals are parsed into shots using one of the shot-boundary detection algorithms, shot information is also used in audio analysis to determine whether the soundtrack of that shot contains speech, silence or music content. This audio classification can be obtained using energy thresholding and frequency analysis [4] and its output will simply be either speech (T) or silence (S) or music (M) for that shot.

Shot segmented video should be further examined to detect faces within frames of each shot. Either a simple skin color detection or a more sophisticated method can be utilize to obtain face (F) or no-face (N) information for that shot. For location analysis, the histograms of consecutive shots can be compared to check whether there is a similarity between the current and the last (also one previous shot before the last) shots using a threshold, probably higher than the threshold used during shot boundary detection. The result of this operation gives whether the observed scene (location) is changed (C) or remained unchanged (U) for consecutive shots.

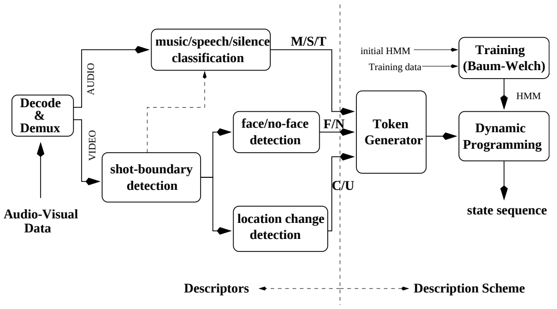


Figure 2: The block diagram of the proposed system

The Token Generator (Fig. 2) simply combines the output of three analysis kernels to obtain one token for each shot. For example, a shot of a movie without a human face and some music in the background will be represented with the token MNC (i.e. Changed location, with No-face and Music at the background). Hence, the token generator will represent each shot with a token consisting of three letters, which are obtained from all possible combinations of sets (S,T,M), (F,N) and (C,U). Finally, the model parameters (state change and state output probabilities) are used to obtain the best state sequence corresponding to the input token sequence by using dynamic programming (i.e. Viterbi algorithm).

4.2. Simulation Results

During our simulations, audio-visual content from MPEG-7 Test Data Set (CD-20 [Spanish TV Movie], 21 [Spanish TV Sitcom], 22 [Portuguese TV Sitcom]) is used for training and test purposes. The ground-truth data is obtained by assigning every shot to a "true" state/scene, subjectively.

The left-to-right and circular topologies are compared according to a criterion which measures the difference between ground truth and computed state sequence output. *Shot accuracy*, R_1, simply finds the ratio of correct assignments to the total number of shots. The results are tabulated in Table 2. It should be noted that for circular topology, the input data (tokens) is pre-segmented such that they contain one establishing scene, one dialogue scene and one transitional scene. Both left-to-right and circular topologies are trained by all the available data. The observation symbols are selected as combinations of audio, face and locations information. A typical result for circular HMM is presented at Figure 3 with some randomly selected keyframes for each scene.

Table 2: *Shot accuracy*, R_1, for different test data sets

R_1	CD-20	CD-21	CD-22
Left-to-right	0.92	0.98	0.99
Circular	0.71	0.82	0.94

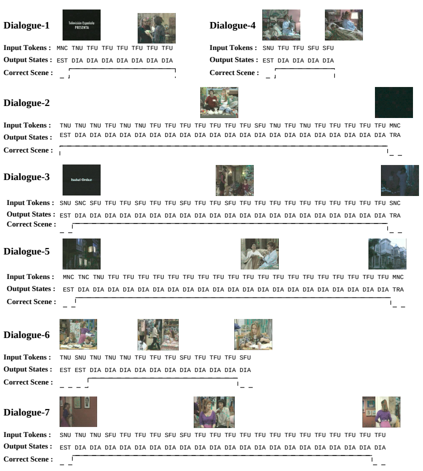

Dialogue-1
Televisión Española PRESENTA
Input Tokens : MNC TNU TFU TFU TFU TFU TFU TFU
Output States : EST DIA DIA DIA DIA DIA DIA DIA
Correct Scene :
Dialogue-4
Input Tokens : SNU TFU TFU SFU SFU
Output States : EST DIA DIA DIA DIA
Correct Scene :
Dialogue-2
Input Tokens : TNU TNU TNU TFU TNU TNU TFU TFU TFU TFU TFU TFU TFU SFU TNU TFU TNU TFU TFU TFU TFU TFU MNC
Output States : EST DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA TRA
Correct Scene :
Dialogue-3
Isabel Ordaz
Input Tokens : SNU SNC SFU TFU TFU SFU TFU TFU SFU TFU TFU SFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU SNC
Output States : EST DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA TRA
Correct Scene :
Dialogue-5
Input Tokens : MNC TNC TNU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU MNC
Output States : EST DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA TRA
Correct Scene :
Dialogue-6
Input Tokens : TNU SNU TNU TNU TNU TFU TFU TFU SFU TFU TFU TFU SFU
Output States : EST EST DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA
Correct Scene :
Dialogue-7
Input Tokens : SNU TNU TNU SFU TFU TFU TFU SFU SFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU TFU
Output States : EST DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA DIA
Correct Scene :

Figure 3: Dialogue scene analysis for MPEG-7 Test Data Set (CD-21) using a circular HMM (Ground truth is shown as a pulse diagram while dialogues are solid lines).

While both topologies are quite successful for assigning shots to correct scenes, the simulation results reveal that left-to-right topology performs better compared to its circular counterpart. However, left-to-right topology requires a-priori knowledge about every scene, hence usually it is difficult to use in practice. The next stage of simulations are conducted on circular topology to test its performance on different observation symbol and training data sets.

At this stage, the performance between the ground truth and computed values is compared using not only R_1 , but also *scene accuracy*, R_2 parameter, which is defined as the ratio of correct scene assignments to the total number of scenes, either dialogue or non-dialogue. For this parameter, the transition point between scenes does not need to be exact but a disjoint scene is strictly necessary for considering a correct scene detection.

Three different set of observation symbols are tested, as audio-only, audio+face and audio+face+location. These different set of data are also tested for different training data, as self-training by the test data, training by all data including and excluding test data, respectively. The results of this experiments are shown in Fig. 4, after combining the individual results for the available three test data sets, CD-20, 21 and 22.

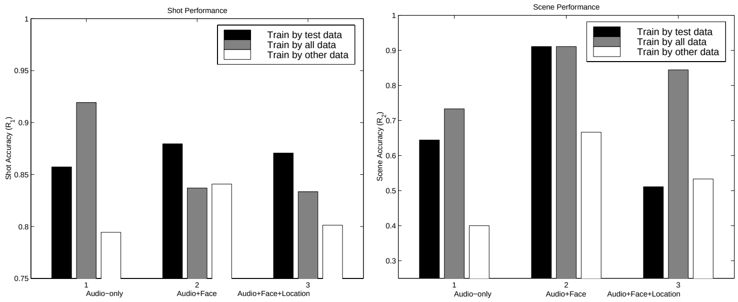


Figure 4: R_1 (left) and R_2 (right) for different observation sets and training data

According to the simulation results, the performance improvement due to utilization location information is not significant. Audio-only observation set still provides good results but its performance depends on training data.

5. CONCLUSIONS

A novel dialogue scene detection scheme is proposed with multi-modal inputs and probabilistic modeling. For both circular and left-to-right HMM topologies, the scene transitions are determined with a good accuracy. For the more practical circular topology, among three observation symbol sets, the set with audio and face information gives the best performance for different training data.

6. REFERENCES

[1] R. M. Bolle, B. -L. Yeo and M. M. Yeung "Video Query : Research Directions," *IBM Journal of Research and Development*, vol. 42, pp. 233–252, 1998. (also avaiable at http://www.almaden.ibm.com/journal/rd/422/bolle.txt).

[2] J. Huang, Z. Liu and Y. Wang "Integration of Audio and Visual Information for Content-based Video Segmentation," in *Proceedings of ICIP'98*, 1998.

[3] S. Tsekeridou and I. Pitas "Speaker Dependent Video Indexing Based on Audio-Visual Interaction," in *Proceedings of ICIP'98*, pp. 358–362, 1998.

[4] C. Saraceno and R. Leonardi "Identification of Story Units in Audio-Visual Sequences by Joint Audio and Video Processing," in *Proceedings of ICIP'98*, pp. 363–367, 1998.

[5] J. Nam, M. Alghoneiemy and A. H. Tewfik "Audio-Visual Content-based Violent Scene Characterization," in *Proceedings of ICIP'98*, pp. 353–357, 1998.

[6] W. Wolf "Hidden Markov Model Parsing of Video Programs," in *Proceedings of ICASSP'97*, pp. 2609–2611, 1997.

[7] J. S. Boreczky and L. D. Wilcox "A Hidden Markov Model Framework for Video Segmentation Audio and Image Features," in *Proceedings of ICASSP'98*, pp. 3741–3744, 1998.

[8] L. R. Rabiner and B-H. Juang. *Fundementals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, USA, 1993.