Two-Level Multi-Modal Framework for News Story Segmentation of Large Video Corpus

Lekha Chaisorn, Chunkeat Koh, Yunlong Zhao, Huaxin Xu, Tat-Seng Chua, and Tian Qi
School of Computing, National University of Singapore, Singapore

1. INTRODUCTION

To tackle the problem of story segmentation, we proposed a two-level multi-modal framework [Chaisorn et al. 2002]. First we analyze the video at the shot level using a variety of low and high-level features, and classify the shots into pre-defined categories using Decision Tree [Quinlan 1986]. Next we perform HMM [Rabiner 1993] analysis in order to identify news story boundaries. This two level framework has been found to be effective in overcoming the data sparseness problem in machine learning. Our approach is similar to the idea of natural language processing (NLP) research in performing part-of-speech tagging at the word level, and higher-level analysis at the phrase and sentence level [Dale 2000].

This paper discussed our enhanced work on performing story segmentation on a large news video corpus used in TRECVID evaluation [TRECVID 2003].

Briefly, the content of this paper is organized as follows. *Section 2* discusses the design of the multi-modal, two-level story segmentation and classification framework. *Section 3* presents the details of shot classification. *Section 4* discusses the details of story segmentation while *Section 5* discusses the experiment results. *Section 6* contains our the conclusion and discussion of future work and *Section 7* presents related work.

2. DESIGN OF THE SYSTEM

We had presented a two-level, multi-modal framework for news story segmentation and classification in our previous work [Chaisorn et al. 2002]. The system works on two levels--shot level and story level. At the shot level, we perform shot classification/tagging which assigns an appropriate shot tag-ID to each of the input shots. At the story level, we employ HMM framework and use the tag-ID, scene/location changed, and speaker changed information to perform story boundary detection. We could achieve the accuracy of about 90%. Because of the data sparseness, our two-level framework had been demonstrated to be superior to the one level framework.

In this paper, we discuss our enhanced system which is scaled to work with the very large data set. Surprisingly, little changes to the original system design is needed. The major enhancements are (a) in shot tagging process, we introduced additional shot categories that appear in CNN and ABC news videos in these corpuses; and (b) in HMM framework, in addition to shot tag-ID and scene/location changed features, we incorporated cue-phrase feature to perform story segmentation. Here, we dropped the use of speaker changes feature as it degrades our system performance. Details of the enhanced system are discussed in the following sections.

2.1 Overview of the System Components

We analyze the raw video using a two-level story segmentation scheme as proposed in Chaisorn et al (2002). The basic unit of analysis is the shots, and we employ multi-modal analysis involving visual, audio and textual features. Briefly, we model each shot using high-level object-based features (face, video text, and shot type), temporal features (background scene change, speaker change, motion, audio type, and shot duration), and low-level visual feature (color histogram). At the shot level, we employ the Decision Tree to classify the shots into one of predefined genre types (details are discussed in Section 2.2). We then perform HMM analysis to detect story boundaries using the shot genre information, as well as time-dependent features based on speaker change, scene change and cue-phrases. The overall story segmentation scheme is shown in Figure 1.

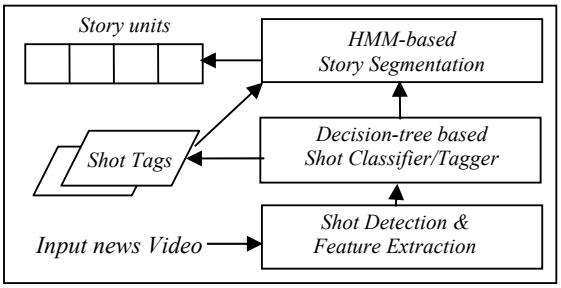

Figure 1: **Overall system components**

In addition, in this paper, we discussed the method of classification of news story. Heuristic rule-based technique was adopted to further classify the detected stories from our two-level approach into "news" or "miscellaneous". Details of shot tagging/classification and story segmentation/classification will be discussed in Section 3 and 4 respectively.

2.2 Selection of Shot Categories

This step is to determine an appropriate and complete set of categories to cover all shot types for these corpuses The categories must be meaningful so that the category tag assigned to each shot is reflective of its content and facilitates the subsequent stage of segmenting and classifying news stories. We studied the set of categories employed in related works and the structures of typical news video. We arrived at the following set of 13 shot categories*: Intro/Highlight, Anchor, 2Anchor, Meeting/Gathering, Speech/Interview, Live-reporting, Still-image, Sports, Text-scene, Special, Finance, Weather, and Commercial* as proposed in the previous paper. In addition to these categories, we introduced additional categories to capture the specific shots used frequently in TRECVID videos, i.e. CNN and ABC news. The five new categories are *"LEDS*": to represent lead-in/out shots; "TOP": to model top story logo shots; "SPORT": to capture sport logo shots; "PLAY": to represent play of the day logo shots; and "health": to model health logo shots. We dropped the use of "still-image shots because it is often that there is moving text at the bottom of the frame, our algorithm which used motion feature, failed to detect this type of category. Thus, the total number of shot categories is 17 which cover all essential types of shots in this collection. Some categories are quite specific such as the Anchor or Speech categories. Others are more general like the Sports or Live-reporting categories.

For completeness, we also subdivided the sports story into sub-stories depending on different types of sports. This is also a requirement of TRECVID for story segmentation task. Figure 2 shows examples of shot categories in our framework.

Figure 2: **Examples of shot categories**

2.3 The Selection of Features

We selected the features that can be automatically extracted and are essential to differentiate one class from the others. These features are:

a. Color Histogram: It models the visual composition of the shot, and is particularly useful to resolve several scenarios in shot classification. This feature is used in the detection of "Weather", "Finance", "Anchor", "2anchor", "TOP", "SPORT", "LEDS", "PLAY", and "health" shots.

b. Scene change: This feature indicates whether there is a change of scene between the previous and current shots. It is derived by computing the difference in color histograms of key frames between the current and previous shots.

***c Audio*:** This feature is very important especially for Sport and Intro/Highlight shots. For Sport shots, its audio track includes both commentary and background noise, and for Intro/Highlight shots, all the narrative is accompanied by background music

e. Motion activity: We classify the motion into *low* (like in an Speech/Interview shot where only the head region has some movements), *medium* (such as those shots with people walking), *high* (like in sports), or *no* motion (for still frame or Text-scene shots).

f. Shot duration: This feature was employed in both shot classification and news story classification. It helps to resolve the ambiguities between "news" and "misc'" stories.

***g. Face*:** We extract in each shot the number of faces detected as well as their sizes. Shots with one or two faces detected are further differentiated into Anchor, 2Anchor, or other shots. The size of the face is used to estimate the shot types.

***h. Shot type*:** We divide the shot type into *closed-up, medium-distance* or *long-distance* shot based on the size of the face detected in the frame.

i. Videotext: A text-scene shot typically contains multiple lines of centralized text such as the results of a soccer game. Hence, for each shot, we simply extract the number of lines of text appear in the key frame and determine whether the text is centralized

j Cue-phrase: We have included cue-phrase feature, i.e. for each shot, we determine whether there is a presence of cue-phrase at the beginning of the shot.

3. THE CLASSIFICATION OF SHOTS

News is a rather structured media with regular structures. It consists of a wide variety of shot types arranged in a well-defined sequence designed to convey the information clearly to a wide range of audiences. Certain shot types like commercials, studio anchor person shots, finance and weather shots etc, have well-defined and rather fixed temporal-visual characteristics. They can best be detected using specific detectors. For the rest of the categories, a learning based approach using Decision Tree is used for their classification.

3.1 Shot Classification Process

3.1.1 Commercial detection

Commercial blocks and individual commercials are usually preceded and ended with a sequence of black frames. Also, the ASR recognition rate during the commercials is usually low, as there is more background music/noise. Hence, commercials tend not to have any recognized ASR outputs. The process of commercials detection is therefore accomplished in the following two steps; a) black frames detection using color histogram; and b) commercials blocks detection using clustering technique and a combination of black frames, silence and low ASR confidence level.

3.1.2 Identifying Anchor and 2Anchor shots

For these corpuses, we observed that, anchor persons always appear in three different positions, i.e. left, center, or right position. Thus, in order to eliminate those face detected shots that unlikely to be *Anchor* shots, we used the number of faces detected from the key frame of each shot and their positions to identify the *Anchor* and *2Anchor* shots. For shots with one face detected, we use the size of the detected face and its position to classify the face into one of the three different types.

For shots where the detected face satisfies our thresholds for position and size, we extracted their LUV color histogram and performed clustering using the single-link clustering algorithm. Since the number of clusters needed to obtain optimum result varies from video to video, we processed the frames for each video starting with 2 clusters and increasing the number of clusters by one, until the largest cluster contains less than or equals to 24 shots (average number of anchor shots for one video in the development set). The cluster with the largest number of shots will be the *Anchor/2Anchor* shots. Finally, we separate the *Anchor* from *2Anchor* shots by detecting the number of faces.

3.1.3 Visual-based shot detection

Visual-based shots are the shots that depend on broadcast station produced to present their programs. These programs are regularly aired in a certain period of time within a broadcast news. In these corpuses, these visual-based shot categories are, "Finance", "Weather", LEDS, "health" logo, "SPORT" logo, and "TOP" (Top stories) logo. We used 176 Luv color histogram as the feature, and employ image matching and video

sequencing techniques developed in our lab to perform the detection.

3.1.4 Rule-based Shots Detector using Decision Tree

The remaining shots were classified using Decision Tree. The feature vector of a shot is of the form:

$$S_i = (a, m, d, f, s, t, c) \quad (1)$$

Where *a* is the class of audio, $a \in$ {t=speech, m=music, s=silence, n =noise, tn = speech + noise, tm= speech + music, mn=music+noise}; *m is* the motion activity level, $m \in$ {l=low, m=medium, h=high}; *d is* the shot duration, $d \in$ {s=short, m=medium, l=long}; *f is* the number of faces, $f \geq 0$; *s is* the shot type, $s \in$ {c= closed-up, m=medium, l=long, u=unknown}; *t is* the number of lines of text in the scene, $t \geq 0$; *and c is* set to "true" if the videotexts present are centralized, c $\in$ {t=true, f=false}.

4. STORY SEGMENTATION AND CLASSIFICATION

As the requirements from the TRECVID, we have to perform story segmentation based on different set of features. First, only video and audio features can be used. Second, the segmentation is based on the given ASR output, and third, the segmentation that uses the combination of video, audio and the ASR features.

4.1 The segmentation Using Video-Audio based features

After the shots have been classified into one of the pre-defined categories, we employ the HMM technique to detect story boundaries. We use the shot sequencing information, and examine both the tagged category and appropriate features of the shots to perform the analysis. We represent each shot by: (a) its tagged category; (b) scene/location change (1= change, 0 = unchanged), and (c) cue-phrase (1=present of cue-phrase, 0= no cue-phrase).

$$S = [t, l, c] \quad (2)$$

where '*t'* is the tag-ID of a shot*;* '*l'* is the scene/location change feature, and '*c'* is the cue-phrase feature of the shot respectively. From Equation (2), it can be seen that each output symbol is represented by 1 of 17 possible categories, 1 of 2 possible scene/location changed feature, and 1 of 2 cue-phrase feature. This gives a total of 17x2x2 = 68 distinct vectors for modeling using the HMM framework. For more details on our HMM framework, refer to our paper [Chaisorn et al. 2002].

4.2 The segmentation using ASR based features

We divided the task under text segmentation using the ASR result given by TRECVID into four main tasks. They are multi-Resolution Analysis (MRA), cue-phrase detection, commercial block detection, and news classification. Figure 3 depicts the system processes.

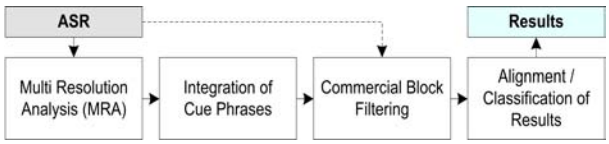

Figure 3: Processes in ASR-based segmentation

4.2.1 Multi-Resolution Analysis (MRA)

For text-based segmentation of the video, we make use of the multi-resolution analysis and wavelet transformation techniques in [Li Yang, 2001]. We adopted the term-based and domain independent approach, which relies only on word variations across segments of text to detect topic change. The process for text segmentation can be explained in the following steps:

1. First, stop words such as "the", "a" are removed from the transcript and the remaining terms are stemmed using Porter's stemming algorithm. Each term *k* at position *i* of the transcript is constructed as the unit term vector:

$$T(i) = t_k = [f_{k_1}, f_{k_2}, ..., f_{k_M}], \quad (3)$$

where $f_{k_j} = \begin{cases} 1 & \text{if term } j = \text{k} \\ 0 & \text{otherwise} \end{cases}$

M = number of unique terms in the transcript

And the transcript *D* is represented as this sequence of term vectors:

$$D = [T(1), T(2), \ldots., T(N)], \quad (4)$$

where N = number of non-trivial stemmed terms in *D*

In our tests, the average values of M = 1000, N = 2500.

2. By using Canny wavelets of different resolution, different window sizes of text corresponding to the span of the wavelet can be modeled and analyzed.
3. For each point *k* in the transcript, we perform the convolution of the wavelet and T(*x*) within the window ±z/2 from k to obtain the wavelet value $W_{a,k}f(x)$. This value measures the Euclidean distance between the left and right segments of text within the window:

$$W_K = EuDist(D_K^L, D_K^R) = \sqrt{\sum (d_{K_i}^L - d_{K_i}^R)^2} \quad (5)$$

Next we generate a graph of W_K. The peaks correspond to topic transitions detected within the document. For more details of this method, refer to [Liyang 2001].

4.2.2 Cue Phrase Detection

There are two types of cue-phrases that we need to identify, *Begin/End Cue-Phrases* and *MISC* cue-phrases.

a) *Begin/End Cue-Phrases:* We identified that some of the story segments are typically preceded or ended with a set of cue-phrases. To extract this list of cue-phrases, we first compile a list of unique n-grams from the ASR transcript in all the story segments from the development data set. For each n-gram t_i, we calculate, p_b, the probability that the n-gram indicates the start and p_e, the probability that indicates end of the story.

The list of p_b and p_e are ranked, and we select the top n-grams with $p(t_i) \geq 0.80$ as the cue-phrases. We consider first and last 10n-gram in each story. Examples of begin-cue-phrases in these corpuses are "checking the hour's"; "good evening i'm"; and examples of end-cue-phrases are "abc news Washington"; "cnn new york", etc..

b) MISC cue-phrases

In addition, miscellaneous (*misc*) segments contain similar information such as reporter chit-chat/scoreboard/stock quotes/advertisements as defined in the TDT-2 guidelines. Using similar method, we first obtain the list of unique n-grams from the *misc* segments, and for each n-gram t_i, we compute the

probability: $P_{misc}(t)$ equals to the number of *MISC* stories containing n-gram t_i divided by the total number of stories contain ti. After we obtained $P_{misc}(t)$, next, the top ranked n-grams are selected and clustered to generate a list of *misc* cue-phrases. Examples of *misc* cue-phrases are "Weather forecast is next", "when we come back", "on the score board", etc.

4.2.3 Commercial Block Detection and Filtering

a) As commercial blocks tend to contain many incorrectly transcribed word tokens and irrelevant information, they are detected and filtered in the segmentation process. From the list of informative features that the LIMSI group has provided in the ASR transcript, a commercial block classifier is implemented based on: (a) c*ommercial timing information*; (b) *long silence duration*; (c) *low Averaged ASR confidence;* and (d) *preceding cue-phrases*

4.2.4 Alignment and Classification of Results

From the ASR of the development set, we found that 96% of the story boundaries are located at silence intervals ≥ 0.2 seconds. We incorporate this knowledge by aligning the results from MRA to the closest silence or speaker change using the distance measure:

$$D(y,x) = \frac{\alpha_s}{|y-x|} SilenceDur(y) + \alpha_c SpkrChange(y) \qquad (6)$$

where y: potential boundary; x: detected boundary from MRA; α_S, α_c: arbitrary weights; SpkrChange(y): 1 if speaker change at y, 0 otherwise.

For the classification of results, segments in the video are classified as *misc* if it is detected as a commercial block or contains *misc* cue-phrases. The remaining segments are labeled as *news*.

4.3 The segmentation using a combination of video-audio and ASR based feature.

From the result of the segmentation based on video and audio, we perform the alignment using the misc cue-phrases as described in *Section 4.2.* After the alignment, we will obtain the more precise story boundaries.

5. TESTING AND RESULTS

5.1 Training and Test Data

The training and test data are CNN and ABC news of the year 1998. Altogether, there are about 120 hours (240 videos, each about half an hour in duration), 112 videos are used for the development set, and the remaining is used for the test set.

5.2 Shot Level Classification

In our previous paper, we used one day of news (half an hour) for training and another day for testing; we could obtain the shot classification accuracy of 95%. Here, we tested the shot classification on 20 videos, 10 each from CNN and ABC, our initial result shows that we can obtain the accuracy about 85%. The accuracy is lower than that of our previous paper because, here there are more categories and more techniques has been incorporated. Moreover, the data set is much larger. Most of the errors are from the detection of those temporal-visual based shot types, for example "LEDS", "TOP", etc. These types of shot appear in a very short duration, thus our algorithm which is designed to handle larger videos failed to detect them effectively.

5.3 News Story Segmentation and Classification

5.3.1 News Story Segmentation

We set up five experiments, 1, 2, 3, 4, and 5 required. For experiment 1 and 3, we used tag_ID and scene/location change as the features. As for experiments 2 and 4, we used the same features as the experiments 1 and 3, but also included cue-phrase feature. These four experiments, we employed HMM framework as described earlier to locate story boundaries. We performed initial experiments by varying the number of states from 4 to 15 to evaluate the results. From our training and test data set (both selected from the development set), our initial test indicates that the number of state equals to 11 gives the best result for experiment 1 and 3, and the number of state equals to 13 gives the best result for experiments 2 and 4. As for experiment 5, we perform story segmentation using the ASR based feature as described in *Section 4.2*. The experimental results evaluated by TRECVID are presented in table 1.

Table 1: Presents our story segmentation results and TRECVID evaluation.

Exp	T	Total BD	SubBD	Found inSub	Found inTruth	Re (%)	Pr (%)
1	1	2929	2919	2156	2105	71.87	73.86
2	2	2929	2825	2199	2158	73.68	77.84
3	1	2929	2812	2132	2084	71.15	75.82
4	2	2929	2731	2166	2127	72.62	79.31
5	3	2929	2433	1402	1383	47.22	57.62

Note: T – type (1=Video+Audio, 2=Video+Audio+ASR, 3=ASR), Total BD – total reference boundaries in truth data, SubBD – Submitted boundaries, FoundinSub – Boundaries found in our submitted result, FoundinTruth – Reference boundaries found in the truth data, Re – recall (FoundInTruth/TotalBD), Pr – precision (foundInSub/SubBD)

5.3.2 News Classification

For the first run (only video and audio features are allowed), we introduced heuristic rule-based approach to classify the detected stories into "news" or "misc". For the first shot of each detected story, we identify its category. This category was obtained during the shot tagging process as discussed in Section x. The category gives us some cue whether the detected story is likely to be "news". For example, if the first shot is *Anchor* shot, it is likely that this story is considered as "news". However, it is not always true. For instance, the story that begins with *Anchor* shot in which the anchor person is introducing upcoming news after the commercials. This story is considered as "misc". In this case, we need the shot category information of the current and successive stories. Furthermore, story duration is also important to differentiate an ambiguity between "news' and "misc". Therefore, in order to perform the classification effectively, we also need the shot category information of the successive stories as well as the current story duration. For the second run (in addition to video and audio, ASR features is included), we used the result from the first run and performed the alignment based on the miscellaneous cue-phrases.

The algorithms/rules for story classification are given below:

a) The Common rules for both ABC and CNN news

```
rule 1. if (Curr = COMMERCIAL),     then the story is  "misc"
rule 2. if (Curr = LEDS),     then the story is  "misc";
rule 3: if (Curr = Intro/Highlight), then the story is  "misc";
rule 4. if (Curr = ANCHOR) and (Next = LEDS) and
   story duration <=TOLERANCE,  then the story is  "misc";
rule 5. if (Curr = ANCHOR) and (Next = COMMERCIAL)
            then the story is  "misc";
rule 6. if (Curr = ANCHOR)
            if story_dur <=TOLERANCE), then the story is
            "misc" ,     else the story is  "news";
rule 7. if (Curr =  2ANCHOR) and (story duration <=
            TOLERANCE), then the story is  "misc";
rule 8. if (Curr = OTHERS),  then the story is  "news";
```

b) The specific rules for CNN news

```
rule 1: if (Curr = ANCHOR) and ((Next = WEATHER)
            or  (Next = HEALTH) or  (Next = 2ANCHOR) or
            (Next = Intro/Highlight)), then the story is  "misc";
rule 2: if  (Curr = SPORT), then the story is  "news";
rule 3: if (Curr = WEATHER), then the story is  "news";
rule 4 if (Curr = HEALTH) and (Next  = HEALTH)
            then the story is  "news";
rule 5: if (Curr = TEXT-SCENE) and (Prev = sport)
            then the story is  "misc";
```

Note: In both algorithms a) and b), **Curr -** first shot of the current story**, Next -** first shot of the next story, **Prev** – first shot of the preceded story. TOLERANCE – duration in seconds.

The classification for the third run (ASR based) was discussed in *Section 4.2.4*. The classification results for all runs are present in table 2.

Table 2: The result of news classification

Run	T	News Recall (%)	News Precision (%)
1	1	93.60	93.61
2	2	92.36	96.02
3	1	91.78	95.14
4	2	91.57	96.26
5	3	92.21	77.20

Note: we submitted five runs, Runs 1 and 3 are based on video-audio while Run 2 and 4 are based on video-audio and ASR. The last result, Run5 is based on ASR only.

5.4 News Story Segmentation based on the ASR

For the classification results, we could achieve the accuracy of 93.6% and 93.6 % for recall and precision respectively.

In our previous paper, we could achieve the accuracy for the story segmentation about 90%. From Table 1, the accuracy from experiment A (using video and audio based features) is lower than that of our previous paper because of several reasons. First, according to TRECVID guidelines, each submitted boundary must lie within the tolerance of 5 seconds (in both directions) of the reference boundary. That is, each submitted boundary is allowed up to 5 seconds late or early than the reference boundary. Second, by using only visual-based cue is not sufficient to locate and classify certain detected stories into "misc". For example, the score summarizing scene which normally appears at the end of each sport reporting, this portion is considered "misc". In general, our algorithm detects the whole chunk of sport including these scenes summarizing the scores as one detected story. Third, there are some miscellaneous words that although appear in news story but this portion of news is considered as "misc". For example, "I am <person name> CNN Headlines news" which appears in *Anchor* shots, this duration of the above phrase is classified as "misc". In order to tackle this problem, only text segmentation and classification can do the job. Fourth, the test data set in these corpuses are much larger than our test data in the previous paper. There are other guidelines that if we use only visual cues (video and audio), will not be sufficient to perform the story segmentation and classification adequately. Thus, in experiment B (based on the result from experiment A, plus the use of text feature), we could improve our system performance in both recall and precision as can be seen in table 1.

6. CONCLUSION AND FUTURE WORK

We have presented our framework to perform news story segmentation and classification on large scale data of about 120 hours of video. The system is divided into three layers. They are feature layer, shot layer and story layer. Our framework on story segmentation is a two-stage process, shot level and story level. At the shot level, we employed the Decision trees and several techniques to detect/classify the input video shots into one of the predefined categories, using a combination of features include low level feature such as color, temporal features such as audio type, motion and shot duration, and high level features, such as face/s and video-texts. At the story level, in addition to shot tag-ID (obtained from the shot classification process) and scene changed feature, we also incorporated cue-phrase as the feature to represent each shot. HMM was employed to perform story boundary detection and simple rule based technique was used to classify each detected story into "news story' or "miscellaneous". From the evaluation from TRECVID 2003, for story segmentation, we could achieve the accuracy of ***72.62%*** and ***79.34%*** for recall and precision respectively. As for news classification based on the result of the story segmentation, we could achieve the accuracy of ***93.60%*** and ***93.61%*** for recall and precision respectively. It can be seen that we could obtain high accuracies for both story segmentation and classification. Therefore, it is demonstrated that, our two-level multi-modal framework is very efficient.

Our future work, we are looking at higher order statistical techniques such as the hierarchical HMM to perform news story segmentation.

7. RELATED WORK

In our framework, the works are related to two areas of research; shot classification, and scene transition detection. Researches on shot classification had been reported in [Ide et al 1998][Zhou et al. 2000][Chen and Wong 2001]. Our approach on shot categories is similar to that of Ide et al. We adopted a subset of their defined categories such as Anchor, Speech/report, and gathering. However, we introduced another 13 categories to cover all types of shot in these corpuses. Next, we employed a machine learning based approach, in particular, Decision trees and several techniques, to classify the input video shots into one of the 16 predefined types.

In the area of scene transition detection, most existing techniques incorporate information within and between video segments to determine class transition boundaries using mostly the HMM approaches. Eickeler et al.[1997] employed the HMM

to classify the video sequence into the classes of Studio Speaker, Report, Weather Forecast, Begin, End, and the editing effect classes. Huang et al.[1999] employed the HMM to classified the TV programs into the categories of news report, weather forecast, commercials, basketball games, and football games. Alatan et al. [2001] aimed to detect dialog and its transitions in fiction entertainment type videos. They used HMM to locate the transition boundary between the classes of Establishing, Dialogue, Transition, and Non-dialogue.

Research on news story segmentation that is similar to our work was reported in Hsu and Chang [2003], They used acoustic, speaker identification, face, motion, video-texts, combinations faces and speech, and cue-phrases as the features. They employed maximum entropy based approach to select the features, and used dynamic programming to perform the story segmentation.

In our story segmentation process, we applied the idea from the work of Alatan et al. but we try to locate story boundaries rather the transition between classes. Moreover, we perform the segmentation in two stages, shot level and story level, similar to the approach used in NLP that performs in word level and then sentence level. In addition to shot tag-ID obtained from the Decision Trees and scene changed feature, we have also incorporated cue-phrase as a feature for each shot. We dropped the use of speaker changed feature because it degrades the system performance.

ACKNOWLEDMENTS

The authors would like to acknowledge the support of the National Science & Technology Board and the Ministry of Education of Singapore for the provision of a research grant RP3960681 under which this research is carried out. The author would also like to acknowledge the support I2R for the support of funding and for the helps in many ways. Last, the authors would like to thanks Feng Huamin, Lee Chee Wei, Liu Bin, and Hung Wendong for their helps through out this research.

REFERENCES

[1] A. Aydin Alatan, Alin N. Akansu, and Wayne Wolf (2001). "Multi-modal Dialog Scene Detection using Hidden Markov Models for Content-based Multi-media Indexing". *Multimedia Tools and applications*, 14, pp 137-151

[2] Lekha Chaisorn, Tat-Seng Chua, and Chin-Hui Lee (2002). "The Segmentation of News Video into Story Units", Proceeding of IEEE Int'l Conference on Multimedia and Expo-ICME 2002, Switzerland, Aug 26-29, 2002

[3] Y. Chen and E. K. Wong (2001). " A knowledge-based Approach to Video Content Classification", Proceeding of SPIE Vol. 4315, pp.292-300.

[4] Tat-Seng Chua and Chunxin Chu (1998). Color-based Pseudo-object for image retrieval with relevance feedback. International Conference on Advanced Multimedia Content Processing '98. Osaka, Japan, Nov. 148-162.

[5] Tat-Seng Chua, Yunlong Zhao and Mohan S. Kankanhalli (2000). "An Automated Compressed-Domain Face Detection Method for Video Stratification", Proceedings of Multimedia Modeling (MMM'2000), USA, Nov, World Scientific, pp 333-347.

[6] Robert Dale, Hermann Moisl, and Harold Somers (2000). "Handbook of natural language processing", Imprint New York: Marcel Dekker.

[7] Stefan Eickeler, Andreas Kosmala, Gerhard Rigoll (1997). "A New Approach To Content-based Video Indexing Using Hidden Markov Models", IEEE workshop on Image Analysis for Multimedia Interactive Service (WIAMIS), pp 149-154.

[8] Ichiro Ide, Koji Yamamoto, and Hidehiko Tanaka (1998). "Automatic Video Indexing Based on Shot Classification", Conference on Advanced Multimedia Content Processing (AMCP'98), Osaka, Japan. S. Nishio, F. Kishino (eds), Lecture Notes in Computer Science Vol.1554, pp 87-102.

[9] J. Huang, Z. Liu, Y. Wang (1999). "Integration of Multimodal Features for Video Scene Classification Based on HMM", IEEE signal processing Society workshop on Multimedia Signal processing, Denmark, pp 53-58.

[10] Winston H. –M. Hsu and Shih-Fu Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation", Proceedings of ICME 2003, Baltimore, USA.

[11] Yang Li and Tat-Seng Chua (2001). "Multi-Resolution Analysis on Text Segmentation", Master degree thesis, School of Computing, National University of Singapore.

[12] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Effelsberg. " Scene determination Based on Video and Audio Features", Proceeding of the IEEE conference on Multimedia Computing and System, Volume I, 1998, pp 59-81

[13] J. R. Quinlan, "Induction of Decision Trees. Machine Learning", 1986, vol. 1, pp. 81-106.

[14] L. Rabiner and B. Juang (1993). "Fundamentals of Speech Recognition", Prentice-Hall.

[15] Yi Zhang and Tat-Seng Chua (2000). "Detection of Text Captions in Compressed domain Video". Proceedings of ACM Multimedia'2000 Workshops (Multimedia Information Retrieval), California, USA. Nov, pp 201-204.

[16] WenSheng Zhou, Asha Vellaikal, and C–C Jay Kuo (2000). "Rule-based Classification System for basketball video indexing", Proceedings of ACM Multimedia'2000 Workshops (Multimedia Information Retrieval), California, USA. Nov, pp 213-216.