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

Lekha Chaisorn[+], Tat-Seng Chua[+], Chun-Keat Koh[+], Yunlong Zhao[+], Huaxin Xu[+], Huamin Feng[+] and Qi Tian*

[+] School of Computing, National University of Singapore
* Institute for Infocomm Research, Singapore

ABSTRACT

This paper presents an enhanced work from our previous paper [Chaisorn et al. 2002]. The system is enhanced to perform news story segmentation on a large video corpus used in TRECVID 2003 evaluation. We use a combination of features include visual-based features such as color, object-based features such as face, video-text, temporal features such as audio and motion, and semantic feature such as cue-phrases. We employ Decision Tree and specific detectors to perform shot classification/tagging. We use the shot category information along with two temporal features to identify story boundaries using HMM (Hidden Markov Models). A heuristic rules-based technique is applied to classify each detected story into "news" or "misc".

1. INTRODUCTION

Large amount of broadcast news videos are available. We need an automatic and effective tool to segment these news videos into single-story units. These story units are then used for indexing to support further browsing and retrieval by the users. This paper discusses our enhanced work on performing story segmentation on a large news video corpus used in TRECVID 2003 evaluation. In our approach, we divide the story segmentation process into two levels: shot level that performs shot classification/tagging, and story level that performs story segmentation using HMM framework. This 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]. The results from TRECVID evaluation show that we could achieve the best accuracy of F_1 value of more than ***77%*** for story segmentation and of more than ***94%*** for news classification. Our system outperforms other systems participated in this task.

Our work is related to that of [Christel et al 2002] and [Hsu and Chang 2003]. In particular, Hsu and Chang used similar set of features but performed story segmentation directly using maximum entropy technique.

This paper is organized as follows. Section 2 presents the overview of the system and the choice of news categories and features. Sections 3 and 4 respectively detail the shot classification, and story segmentation and classification. Section 5 describes our experimental results. Finally, we conclude our paper in Section 6.

2. OVERVIEW OF THE SYSTEM COMPONENTS

The key design consideration of our story segmentation framework is in devising a 2-level scheme to analyze the video contents using multi-modal features [Chaisorn et al. 2002]. The use of 2-level scheme helps to alleviate the data sparseness problem in statistical learning. The two levels are: shot classification level, and story segmentation level. The basic unit of analysis is the shots, and we model each shot using a combination of 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 features (color histogram). At the shot level, we employ the Decision Tree to classify the shots into one of the predefined genre types. At the story level, we perform HMM analysis to detect story boundaries using the shot genre information, as well as time-dependent features based on scene change and cue-phrases. The overall story segmentation scheme is shown in Figure 1. In addition, we employ a heuristic rule-based technique to classify the detected stories into the classes of "news" and "miscellaneous".

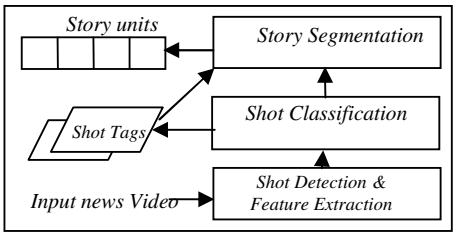

Figure 1: **Overall system components**

The key to the success of the framework is the judicious choice of news categories and features to be used in the shot classification and story segmentation processes. They are discussed in the following sub-sections.

2.1 Shot Categories

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. To achieve this, we use the class taxonomy of TV Any-Time model as the guide. Figure 2 illustrated a general news structure. In addition, we studied the structures of typical news video and the set of categories employed in related work [Ide et al. 1998]. We arrived at the following set of 12 shot categories*: Intro/Highlight, Anchor, 2Anchor, People, Speech/Interview, Live-reporting, Sports, Text-scene, Special, Finance, Weather, and Commercial* as proposed in our previous paper [Chaisorn et al.2002]. 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 additional categories introduced 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. Thus, the total number of shot categories is 17, which cover all essential types of shots in this collection. The shot categories are used in HMM analysis for story segmentation.

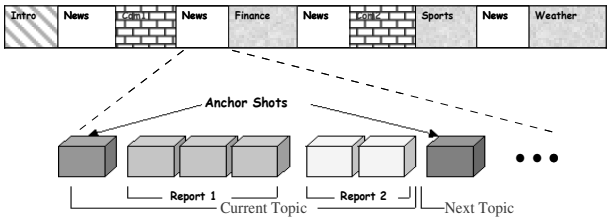

Figure 2: General news structure

2.2 The Selection of Features

In order to support the tasks of shot classification and subsequent story segmentation, we selected the following set of features that are essential to differentiate one class from the others. The other consideration in selecting these features is that they can be automatically extracted using existing tools. The features are:

a) Low level feature

- *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 visual similarity shots like "Weather", "Finance", etc.

b) Temporal features

- *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.
- *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
- *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).
- *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.

c) High level object-based features

- *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 (shots with 2 anchor persons), or other shots. The size of the face is used to estimate the shot types.
- *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.
- *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 appearing in the key frame and determining whether the text is centralized

- *Cue-phrase*: We include typical cue-phrases that appear at the beginning of the news stories. Thus for each shot, we want to know whether such cue-phrases are present or not.

It was demonstrated in our previous paper that face and audio are the most important features for shot classification. However, all the selected features are essential to achieve high accuracy.

2.3 Cue-phrase Extraction

The extractions of all the audio-visual features described in Section 2.2 are well discussed in recent literatures. This Section discusses our technique in extracting cue phrases to support story segmentation. We extract two types of cue-phrases from the ASR derived from the training video set. We extract those that appear at the beginning of news stories, and those that appear in 'misc' story classes. We first compile a list of unique n-grams from the ASR transcripts in all the story segments. For each n-gram t_i, we calculate: (a) p_b, the probability that the n-gram indicates the start of news stories; and (b) p_{misc}, the probability that indicates it is part of a misc-type. The list of p_b and p_{misc} are ranked, and we select the top n-grams with $p(t_i) \geq 0.80$ as the cue-phrases. Examples of the begin-cue-phrases in the news corpus are "checking the hour's"; "good evening I'm". Examples of the *misc* cue-phrases are "weather forecast is next", "when we come back", "on the score board", etc.

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, 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. The following sub-sections describe the shot clusters and the hierarchical steps used to identify shot categoryies.

3.1 Commercial shot detection

Commercial blocks and individual commercials are usually preceded and ended with a sequence of black frames and audio silence. 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. Therefore, we employ clustering technique and use a combination of black frames, long silence duration, typical timing of commercial block within news video and low ASR confidence level to perform commercials detection. Our commercial detection accuracy for this corpus is about 95%.

3.2 Visual similarity shot detection

After we perform commercial detection, the second step is to identify shots in this cluster. For this cluster, there are two sub types: *a) visually similar shots within the video sequence*; and *b) visually similar shots within the broadcast station*. Examples of the first type are *Anchor* and *2Anchor* shots. Examples of the latter type *are Weather*, *Finance,* etc. For the first type, we perform clustering on the shots with detected face/s. For the second type, we use 176-Luv-color-histogram as the feature, and employ image similarity matching and video sequence matching techniques developed in our lab to perform the detection.

3.3 Rule-based Shot classification

The remaining shots are classified using the Decision Tree. The feature vector used for each shot is of the form:

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

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 found are centralized, c $\in$ {t=true, f=false}.

4. STORY SEGMENTATION AND CLASSIFICATION

As part of the requirements from TRECVID 2003, we are required to perform story segmentation based on different classes of features: The three tasks defined are (i) using only video and audio features; (ii) using only ASR; and (iii) based on the combination of video, audio and ASR features.

4.1 Segmentation Using Audio-Visual Based Features (Task i), and Combination of All Features (Task iii)

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, *t;* (b) scene/location

change, *l* (1= change, 0 = unchanged), and (c) cue-phrase at the beginning of story, *c* (1=present of cue-phrase, 0= no cue-phrase).

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

Note that for Task (i) that uses only video and audio features, the cue phrase feature is not used.
From Equation (2), it can be shown that for Task (iii) that uses the full set of features, each output symbol is represented by 1 of 17 possible categories, 1 of 2 possible scene/location changed feature values, and 1 of 2 cue-phrase feature values. This gives a total of 17x2x2 = 64 distinct vectors for modeling using the HMM framework.

After we perform story segmentation, we introduce heuristic rules to classify the detected stories into the classes of "news" or "misc". For the first shot of each detected story, we identify its category. This category (discussed in sub-Section 2.1) was obtained during the shot tagging process as discussed Section 3. The category gives us the clues on whether the detected story is likely to be "news". For example, if the first shot is an *Anchor* shot, then it is likely that this story is "news". However, this is not always true. For instance, the story that begins with *Anchor* shot in which the anchor person is introducing the upcoming news after the commercials, this story is considered as "misc". In this case, we need to know the shot category information of the first shot of the current and successive stories. Furthermore, story duration is also important to differentiate the 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 runs that incorporate ASR (Runs 2, 4 & 5), we use the miscellaneous cue phrases to realigning the story boundaries.

4.2 Segmentation Using Only ASR Based Features (Task ii)

We employ Multi-resolution Analysis (MRA) [Li 2001] on the ASR text provided by TRECVID to extract story boundaries. We adopt the term-based and domain independent approach, which relies only on word variations across segments of text to detect topic change. More details of this method can be found in [Li 2001]. After we have obtained a set of boundaries based purely on ASR (text) analysis, we need to perform post processing to re-align the boundaries to the correct ones. From the ASR of the development set, we found that 96% of the story boundaries are located at the silence intervals of ≥0.2 seconds. We thus incorporate this knowledge by re-aligning the results from MRA to the closest silence or speaker change using the distance measure:

$$D(y, x) = \frac{\alpha_s}{|y - x|} SilenceDur(y) + \alpha_c SpkrChange(y) \quad (3)$$

Where *y*: potential boundary; x: detected boundary from MRA; α_S, α_c: arbitrary weights; Spkr Change (y): 1 if speaker change at y, 0 otherwise. Finally, we classify the stories as "*misc*" if it is detected as a commercial block or contains '*misc*' cue-phrases. The remaining stories are labeled as "*news*".

5. TESTING AND RESULTS

5.1 Training and Test Data

The training and test data provided by TRECVID 2003 are CNN and ABC news video of the year 1998. Altogether, there are about 120 hours of news video. About 60 hours of the videos, called the development set, is used for training the system; while the rest is for testing.

5.2 Shot Classification

We report our results on shot classification based on a subset of TREC videos. In particular, we tested on 20 videos, 10 each from CNN and ABC. Our initial results show that we could achieve an accuracy of about 85%. The accuracy is lower than that of our previous paper because the test set is much larger. Moreover, there are more categories that need to be incorporated. Our analysis shows that most of the errors are from the detection of those temporal-visual based shot types, for example "LEDS", "TOP", etc. These types of shots typically appear in very short durations, thus our algorithm which is designed to handle longer videos failed to detect them effectively.

5.3 News Story Segmentation and Classification

We set up five runs to test the use of different combination of features for news story segmentation.

- Run 1: Recall-priority run using video-audio feature without ASR, i.e. we use tag-ID and scene/location change features only.
- Run 2: Recall-priority run using video-audio feature plus ASR, i.e. we use tag-ID, scene change and cue phrase features.
- Run 3: Precision-priority run using the same feature set as Run 1.
- Run 4: Precision-priority run using the same feature set as Run 2.
- Run 5: Text run using only the ASR-based features.

For the first four runs, we employed the HMM framework as described in Section 4.1 to locate the story boundaries. We sub-divided the development set into training and validation sets. We performed the initial experiments by varying the number of states from 4 to 15. Our initial results indicate that the number of states equals to 11 gives the best results for Runs 1 and 3, and the number of states equals to 13 gives the best results for Runs 2 and 4. As for Run 5, we perform story segmentation using the ASR based feature only. The experimental results evaluated by TRECVID are presented in Table 1.

Table 1 shows that we could achieve the best recall of *74.9%* and best precision of *80.2%* for the story segmentation task. Moreover, we could achieve the best recall of 93.7% and best precision of 96.5% for news classification. Our system is one of the best-performed systems.

Table 1: Results of story segmentation based on TRECVID 2003 news video corpus

Run	T	Recall	Precision	News Recall	News Precision	F1	F1-news
1	1	0.741	0.746	***0.937***	0.939	0.743	0.938
2	2	0.76	0.787	0.925	0.963	0.773	***0.944***
3	1	0.734	0.766	0.918	0.953	0.750	0.935
4	2	***0.749***	***0.802***	0.916	***0.965***	***0.775***	0.940
5	3	0.488	0.584	0.921	0.773	0.532	0.841

Note: T – type (1 = Video + Audio, 2= Video + Audio + ASR, 3= ASR) Recall –recall for story boundaries, Precision – precision for story boundaries., News Recall – recall for news classification, and News Precision – precision for news classification

5.4 Discussion

In our previous paper, we could achieve an accuracy of about 90% for story segmentation. From Table 1, the accuracy from the experiment using video and audio based features (Tasks i & iii)), is lower than that of our previous studies because of several reasons. First, the test data used here is much larger and more varied than the one we used previously. Second, according to TRECVID 2003 guidelines, each submitted boundary is allowed up to 5 seconds deviation from the reference boundary. This is much stricter than the guideline we used in our previous studies. Third, 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, is considered as "misc". In general, our algorithm detects the whole chunk of sport including these scenes summarizing the scores as one detected story. Fourth, there are some miscellaneous words that may also appear in news stories, thus rendering them being classified as "misc". For example, "I am <person name> CNN Headlines news" appears in both the *Anchor* and "misc" shots. Thus the duration of the above phrase tends to be classified as "misc". In order to tackle this problem, we need to utilize text segmentation and classification. Despite these limitations, our overall system performs the best among all submitted systems.

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 news video. Our framework on story segmentation is a two-stage process, shot level and story level. At the shot level, we used 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 and employed the Decision Tree to classify the input video shots into one of the pre-defined categories. 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' or "misc". It can be seen that we could achieve high accuracies for both story segmentation and classification for large news video corpus. This demonstrates that our two-level multi-modal framework is very effective and our system is one of the best-performed systems in this evaluation. The interface of our news retrieval system is shown in Figure 4.

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

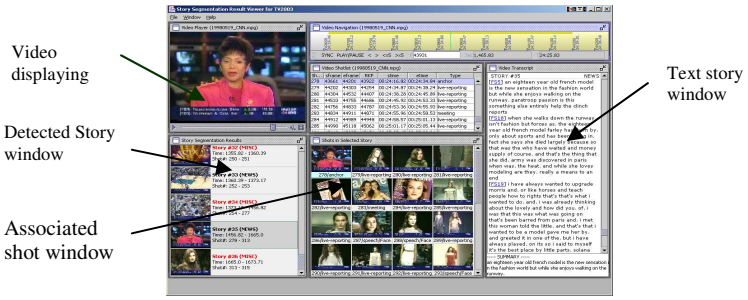

Figure 4: Interface of our System

7. ACKNOWLEDGEMENTS

The authors would like to thanks Lee CheeWei, Liu bin, Hung Wendong, and Wang Ye for their helps through out this research.

8. REFERENCES

[1] L.Chaisorn, T.-S Chua and C.-H.Lee (2002). The segmentation of news video into story units. IEEE Int'l Conf.on Multimedia and Expo. (ICME2002), Lausanne, Switzerland.

[2] W. H.-M. Hsu and S.-F. Chang (2003). A Statistical Framework for Fusing Mid-level Perceptual Features in News Video. Invited paper, ICME 2003, Baltimore, USA.

[3] M.G. Christel, A.G. Hauptmann, H.D. Wactlar and T.D. Ng (2002). Collages as Dynamic Summaries for News Video. In the Proceedings of ACM Multimedia 2002, Juan-les-Pins, France.

[4] 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.

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

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