A Hierarchical Approach to Story Segmentation of Large Broadcast News Video Corpus

Lekha Chaisorn[1], Tat-Seng Chua[1], Chin-Hui Lee[3] and Qi Tian[2]

[1]*School of Computing, National University of Singapore, Singapore*
[2]*Institute for Infocomm Research, Singapore*
[3]*Georgia Institute of Technology, USA*

ABSTRACT

A multi-modal two-level framework for news story segmentation was proposed in Chaisorn et al. [1]. This paper presents our extended work scaled to cope with large news video corpus used in TRECVID 2003 evaluation. We divided our system into two levels: the shot level that classifies input video shots into one of the predefined categories using a hybrid of heuristic and learning based approaches; and story level that performs story segmentation using the HMM framework based on the output of shot level and other temporal features. A heuristic rules-based technique is then employed to classify each detected story into "news" or "misc". We evaluated our system on over 120 hours of news video and showed that our system could achieve an accuracy of more than 77%. Our system came first in the TRECVID 2003 story segmentation task.

1. Introduction

Large amount of information is available for users to access and retrieve on the Internet. One example of such information is the broadcast news videos. News video is informative and rich in contents. However, before a long sequence of news video can be accessed, there is a need to segment these news videos into single-story semantic units. These story units are then used for indexing to support further browsing and retrieval by various users. We proposed a framework to perform automatic story segmentation in Chaisorn et al. [1]. Similar works were done in Christel et al. [2] and Hsu and Chang [3]. This paper discusses our experience in enhancing our framework to perform story segmentation on large news video corpus used in TRECVID 2003 evaluation. The enhancements are: (a) the introduction of additional shot categories to cover new categories found in this collection; (b) the incorporation of cue-phrase feature; (c) the training of a new HMM model based on a large set of data from the development set; (d) the enhancement of techniques to detect anchor/2anchor shots; and (e) the integration of a heuristic rule based technique to classify the detected stories into "news" or "misc". From the TRECVID guidelines, a "news" story is defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses. The rest of coherent segments are labeled as "misc" (miscellaneous). These "misc" stories cover a mixture of footages, including commercials, lead-ins, reporter chit-chat etc. Further details on the guidelines can be found in [6].

Our framework on story segmentation 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.

This paper is organized as follows. Section 2 presents the details of our story segmentation framework. Section 3 discusses experimental results and experiences. Finally, we conclude our paper in Section 4.

2. News Story Segmentation

2.1 Overview of the System Components

The hierarchical structure for our story segmentation scheme is presented in Figure 1.

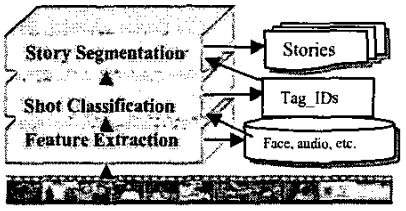

Figure 1: System components

We first extracted features that are essential to shot classification and story segmentation. After the features are extracted, we employed specific detectors and a learning-based approach to perform shot classification. At the story level, we used the shot category information and other temporal features to perform HMM analysis to identify story boundaries.

2.2 Choices of Shot Categories and Features

The categories must be meaningful so that the category tag assigned to each shot is reflective of its content and facilitates the subsequent stages of segmenting and classifying news stories. To achieve this, we used the class taxonomy of TV Any-Time model as the guide, and studied the structures of typical news videos and the set of categories employed in related work

0-7803-8603-5/04/$20.00 ©2004 IEEE

[4]. We arrived at 17 categories of shots that would cover most types of news video. They are: *Intro/Highlight, Anchor, 2Anchor, People, Speech/Interview, Live-reporting, Still-image, Sports, Text-scene, Special, Finance, Weather,* and *Commercials*. In addition to these classes, we introduced *"LEDS"* (to represent lead-in/out shots), *"TOP"* (top story logo shot), *"PLAY"* (for play of the day logo shot), *"SPORT"* (to capture sport logo shots), and *"HEALTH"* (to represent health logo shots). These 17 categories cover all essential types of shots in this corpus.

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. One of the considerations in selecting these features is that they can be automatically extracted using existing tools. Such features include: *(a)* visual-based features like color histogram; *(b)* temporal features such as the audio, motion, and scene-change; and *(c)* high-level features such as face, videotext, and cue-phrases.

Color Histogram: We used 176 Luv color histogram to detect visually similar shots like *"Weather", "Finance"*, etc.

Scene change: This feature indicates whether there is a change of scene between the previous and current shots.

Audio class: For each shot, we classified the audio into the subclasses of speech, silence, music, its combination, or noise. This feature is very important especially for Sports (audio track includes both commentary and background noise) and Intro/Highlight shots (all the narrative is accompanied by background music.).

Motion activity: We classified the motion into low, medium, high or no motion.

Shot duration: We classified shot duration into short, medium, and long. Anchor shots seem to have long duration while live reporting shots are often of very short duration.

Face: We extracted in each shot the number of faces detected as well as their sizes. The size of the face is used to estimate the shot types.

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

Videotext: For each shot, we simply extracted the number of lines of text appear in the key frame and determine whether the texts are centralized.

Cue-phrase: From the ASR (automatic speech recognition) results of the speech track in videos; we analyzed the statistics of cue-phrases that typically appear at the beginning of news stories (Begin-Cue) and miscellaneous stories (Misc-Cue). For each shot, we represented Begin-Cue as 1 (presence of Begin-Cue) or 0 otherwise. This feature is used during story segmentation process. An example of Begin-Cue is "good evening I'm <person name>". An example of Misc-Cue is "when we come back". Misc-Cue is used to align the detected stories.

2.3 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 concisely to a wide range of audiences. A block of commercials can be detected using heuristic technique. Certain shot types like 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 other categories that are non-structural and do not have clearly defined temporal-visual characteristics, we used a learning based approach for their classification.

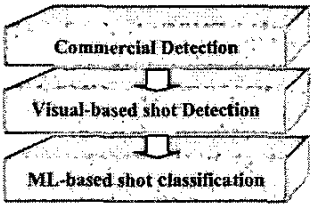

Figure 2: Hierarchical steps in shot classification

In this research, we divided the shots into three main clusters as shown in Figure 2. The first cluster is the *Heuristic-based* cluster for *Commercial* shots. Here, the detection was performed based on a combination of black frames (detected using color histogram), silence and low ASR confident level.

The next cluster is the *Visual-similarity-based* cluster such as *Weather, Finance, Anchor, 2Anchor* shots. For this cluster, we used 176-Luv-color-histogram and employed clustering technique to detect the *Anchor* and *2Anchor* categories on the shots with detected face/s. For the rest of the categories in this cluster, we employed image similarity and video sequence matching developed in our lab to perform the detection.

The last cluster is the *Machine-learning-based* (ML) cluster such as *People, Live-reporting, Text-scene, etc.* shots. Shots of this cluster are classified using Decision Tree. The feature vector used for each shot is of the form:

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

where a is the class of audio, m is the motion activity level; d is the shot duration level, f is the number of faces, $f \geq 0$; s is the shot type; t is the number of lines of text, $t \geq 0$; and c is set to "true" if the videotexts found are centralized, otherwise it is set to "false".

2.4 Story Segmentation and Classification

As part of the requirements from TRECVID 2003 [6], we need to perform story segmentation based using different combination of features. This is to test the

effectiveness of using different features for this task, in particular, the contrast between the text-based ASR features, and the audio-visual (AV) features. The feature sets used are: (*F1*) only AV features; (*F2*) only ASR feature; and (*F3*) a combination of AV and ASR features.

2.4.1 Story Segmentation using Feature Set F1 & F3

After the shots have been classified into one of the pre-defined categories, we employed the HMM approach to detect story boundaries. Figure 3 illustrates our HMM model. We can describe the model using two kinds of state: internal story states; and transition states. Each transition state signals the change to next story while the internal story states indicate that they are within the story.

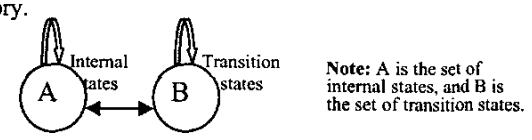

Figure 3: Our HMM model

To perform HMM analysis, we represented each shot by: (a) its tagged category, *t*; (b) scene change, *l* (*l* = c for scene change, and *l* = u otherwise); and (c) Begin-cue, *c* (*c* = 1 for presence of Begin-cue, and *c* = 0 otherwise).

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

From Equation (2), when we use the full set of features (F3), this gives a total of 17x2x2 = 68 distinct vectors for modeling using the HMM framework (17 categories, 2 possible scene changed feature values, 2 possible Begin-cue feature values). More details on HMM analysis is discussed in Section 3.3.

2.4.2 Story Segmentation using Feature Set F2

For text-based segmentation of video, we used the multi-resolution analysis and wavelet transformation technique as described in [5] to identify story boundaries based on the detection of topic changes.

3. Testing and Results

3.1 Training and Test Data

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

3.2 Shot Classification

We tested our shot classification on a subset of development set (~12 hours of video). The result is presented in Table 1. 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.

Table 1: Summary of shot classification results

Cluster	Precision (%)	Recall (%)
Commercials	99.10	95.80
Visual-similarity-based	88.50	93.30
ML-based	91.0	90.0
Average	**92.87**	**93.03**

3.3 News Story Segmentation and Classification

In order to test the performance of our news story segmentation using different feature sets and under different test environments, we conducted five experiments as follows: Runs 1 and 2: recall-priority runs using feature sets F1 and F3 respectively; Runs 3 and 4: precision-priority runs using feature sets F1 and F3 respectively; and, Run 5: using feature set F3, i.e. ASR-based features only. For the first four runs, we employed the HMM framework to locate story boundaries. For each feature set F1 and F3, we trained the system using different numbers of hidden states. From the experiments, the number of hidden states equals to 11 gives the best results for feature set F1 and the number of hidden states equals to 12 gives the best results for feature set F3. Thus, we trained two HMM models to be used in our tests. Table 2 shows an example of how we detect story boundaries when using feature set F3. For Run 5, we performed text segmentation [5] based on the sequence of text from ASR.

Table 2: An example of observed symbols and output state sequence

I_l	1c0	1u0	1u0	2c1	4c0	4u0	6u0	13c0	13u0
O_l	5	3	3	4	1	1	1	7	9

Note: I_l and O_l represent observed symbol sequence and its corresponding output state sequence respectively. In this example, the transition states are states 5, 4, and 7 accordingly.

After story boundaries are identified, we classified the detected stories into the classes of "news" or "misc". We used the assigned tag of the first shot of each story, time constraint, and heuristic rules to perform the classification. The results of both story segmentation and classification for the five runs are presented in Table 3.

Table 3: Results of story segmentation on this corpus

Run	Feature	Story Segmentation			Story Classification		
		Recall	Precision	F1	Recall	Precision	F1
1	F1	0.741	0.746	0.743	0.937	0.939	0.938
2	F3	0.76	0.787	0.773	0.925	0.963	*0.944*
3	F1	0.734	0.766	0.750	0.918	0.953	0.935
4	F3	0.749	0.802	*0.775*	0.916	0.965	0.940

We can see from Table 3 that, we could achieve the best F_1 measures of over 77% for story segmentation, and over 94% for news classification.

Figure 4 compares our results (system name is NUS) based on feature set F3 with other systems participated in TRECVID 2003 evaluations. Our system came out top in the evaluations.

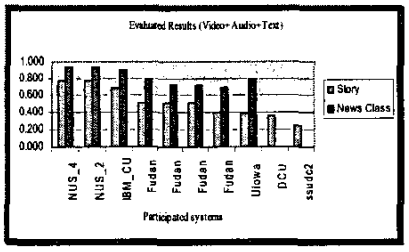

Figure 4: The results compared to the other systems

Further details on the results can be found in [6].

3.4 Discussions

We observed that most of the errors in shot classification and story segmentation were due to errors in feature extraction. Our analysis showed that *face* and *audio classes* are two of the most important features in shot classification, and audio accounts for most of the feature extraction errors.

During the test phase of story segmentation, we found many story patterns that were not discovered by HMM during training. This is because such "unexpected" patterns did not occur sufficiently frequently in training data for HMM to learn the patterns, which is a case of data sparseness problem. One typical story pattern learned by HMM is: Anchor (Tag-ID 2) followed by a remote reporter (Tag-ID 5), and 2Anchor (Tag-ID 3), or the Tag-ID sequence; 2 5 3, which is considered as one story. However, there are patterns of Tag-ID sequence such as: *2 5 3 2 5 3 2 5 3* found during testing that should belong to one story. However, HMM detects that as 3 stories. This leads to over segmentation of stories in many cases. We estimated that this accounts for over 3% of errors in story segmentation. One approach to overcome this problem is to introduce heuristics. But a better solution is to investigate the use of higher order statistics.

Another problem is in segmenting sports news into individual sport news (football, basketball, etc.). This requires specific techniques related to sports which we did not consider. For better performance, there is a need to incorporate better domain specific techniques such as sports classification.

Furthermore, for Misc-cue that appears in the first shot of each detected story, the duration of the above text phrase should be segmented out and classified as "misc". Thus there is a need to re-align the shot boundaries. As our shot segmentation is based on visual features, whereas such re-alignment requires the results of ASR, hence there is a need to integrate both sets of features in shot segmentation. In our results, such re-alignments would help to increase the accuracy by 3-4%.

4. Conclusions and Future Work

We have presented an extension of our 2-level framework to perform news story segmentation and classification on a large news video corpus. Our approach uses a range of features and employs a combination of heuristics, Decision Tree and HMM models to perform story segmentation. Our tests on 120-hour of news video from TRECVID show that our approach is effective, and we could achieve the best F_1 measure of over *77%* for story segmentation and over *94%* for news classification. The interface of our news system is shown in Figure 5.

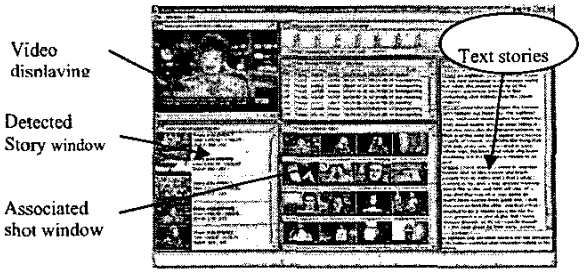

Figure 5: The Interface of our System

The results also point to several directions for future work. One such direction is to explore the use of higher order statistical techniques such as the hierarchical HMM to perform news story segmentation.

4. Acknowledgements

The authors would like to thank Chunkeat Koh, Yunlong Zhao, Cheewei Lee, Huaxin Xu, Huamin Feng, Xiushi Shen, Wendong Huang, Bing Liu, and Ye Wang for their helps in various aspects of this project. The authors would also like to acknowledge the support from Institute for Infocomm research (I^2R).

5. References

[1] L.Chaisorn, T.-S Chua and C.-H Lee, "The Segmentation of News Video into Story Units", *Proceedings of IEEE Int'l Conf. on Multimedia and Expo*, Lausanne, Switzerland, August 2002.

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

[3] W. H.-M. Hsu and S.-F. Chang (2003), "A Statistical Framework for Fusing Mid-level Perceptual Features in News Video", *Proceedings of Intl. Conf. on Multimedia and Expo (ICME)*, Baltimore, USA, July 2003.

[4] I. Ide, K. Yamamoto, and H. Tanaka , "Automatic Video Indexing Based on Shot Classification", *Proceedings of Conference on Advanced Multimedia Content Processing (AMCP)*, Osaka, Japan, 1998.

[5] Y.Li, "Multi-resolution analysis on text segmentation", *Master degree thesis*, School of Computing, National University of Singapore, 2001.

[6] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/tv2003/tv2003.html