Story Boundary Detection in Large Broadcast News Video Archives – Techniques, Experience and Trends

Tat-Seng Chua[1], Shih-Fu Chang[2], Lekha Chaisorn[1] and Winston Hsu[2]
[1]National University of Singapore, Singapore
[2]Columbia University, New York, USA

ABSTRACT
The segmentation of news video into story units is an important step towards effective processing and management of large news video archives. In the story segmentation task in TRECVID 2003, a wide variety of techniques were employed by many research groups to segment over 120-hour of news video. The techniques employed range from simple anchor person detector to sophisticated machine learning models based on HMM and Maximum Entropy (ME) approaches. The general results indicate that the judicious use of multi-modality features coupled with rigorous machine learning models could produce effective solutions. This paper presents the algorithms and experience learned in TRECVID evaluations. It also points the way towards the development of scalable technology to process large news video corpuses.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]:– *Indexing Methods;* H.5.1 [**Multimedia Information Systems**]:– *Video*

General Terms
Algorithms, Performance, Design, Experimentation

Keywords
Story segmentation, news video, machine learning techniques

1. INTRODUCTION
Recent advances in networking, multimedia and WWW have facilitated access to large amounts of multimedia data on the Web. One semi-structured form of information regularly accessed by a large group of users is news, typically in text form but more recently in video form. However, before news video can be conveniently accessed, there is a need to organize the video into meaningful units based on story, which conveys cohesive information about one topic. The story units are then indexed and organized to facilitate retrieval and browsing.

In order to assess the technological progress and spur further research in this important area, TRECVID 2003 [12] has defined news story segmentation as a separate evaluation task. In the 2003 evaluation, over 120 hours of news video are made available and the task is to group sequences of shots into story units using multi-modality features, including closed caption text and automatic speech recognition (ASR) output. According to TRECVID guidelines, a news story is defined as a segment of news broadcast with a coherent news focus containing at least two independent, declarative clauses. The rest of coherent segments are classified as miscellaneous stories, which cover a mixture of footages, including commercials, lead-ins, reporter chit-chats etc. Further details on the guidelines can be found in [12].

MM'04, October 10–16, 2004, New York, New York, USA.

The problem of segmenting news into story units has been actively investigated in text and more recent in video. Typical text-based methods detect coherence in lexical contents and/or discourse structures in text chunks to identify story boundaries [1]. One successful method is based on text tiling, in which the lexical similarity between successive pairs of adjacent text windows is examined, and a story boundary is declared at a location where there is greatest dissimilarity below a threshold [7]. Variants of this method examine coherence in nouns or semantic entities in text [10]. The other category of techniques uses machine learning approach such as the HMM [1], or discourse-based approach [11]. In general, text-based method could achieve an accuracy of around 60% when the correct boundary is defined to be within a certain window. The accuracy is limited as there is insufficient information from text distribution statistics alone to detect accurate boundaries. This accuracy is insufficient for news segmentation. To improve the accuracy, we must utilize multi-modality features, effective structure models and knowledge of news video domain.

This paper reviews recent advances in segmenting news video into story units using text and other modality features, and knowledge of news video production. The review is based mostly on TRECVID 2003 efforts, and thus does not cover efforts outside TRECVID. It presents the algorithms and experience learned in TRECVID evaluations, analyzes the sources of errors, and points the way towards developing scalable techniques for large multimedia contents.

2. NEWS STORY SEGMENTATION TECHNIQUES
This section reviews the techniques employed in recent TRECVID 2003 evaluation of news story segmentation task. The 120-hour of news video provided by TRECVID comes from CNN and ABC for the year of 1998. About 60 hours of the video are used for training and the rest for testing. The video comes with closed caption text and ASR output produced by the LIMSI group [6]. In order to test the techniques over a range of features, participants must submit results that use: (a) text feature based only on ASR output; (b) audio-visual feature; or (c) combination of both types of features. In TRECVID 2003, there were 8 groups submitting a total of 41 runs for the evaluation task.

Figure 1 summarizes the best results in terms of F_1 measure submitted by each group using the above three sets of features on the ~60-hour of test news video. The groups are abbreviated as: NUS [3], IBM/CU [8, 12], UI [4], Fudan [14], DCU [2], SSUDC [10], KDDI [16] and UCF [15]. Caution, however, should be

exercised when using the F_1 measure to interpret the results. This is because it does not provide as complete information as the two-dimensional precision-recall curve. Given a classifier with adjustable tradeoff in the precision/recall space, the F_1 measure tends to fare well when the precision and recall values are close. For example, by changing a threshold for the likelihood ratio in [8], the F_1 values can be improved noticeably.

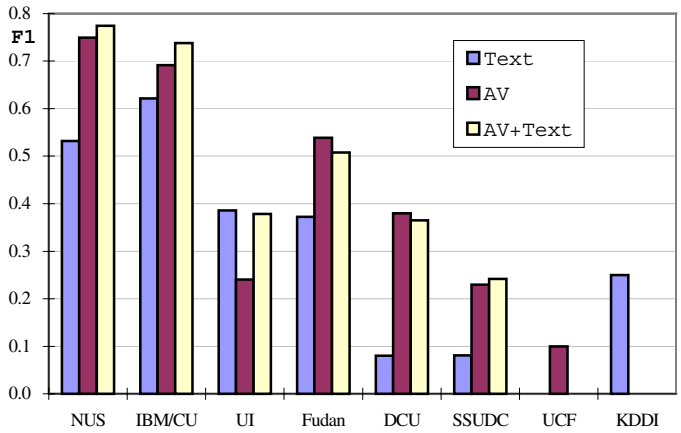

Figure 1: Segmentation results based on text, AV and combination of features

In general, the techniques can be divided into three broad classes. The first class of techniques is text-based, and thus it would inherit the limitations of most text-based methods. The second class is the heuristic rule-based methods that exploit the specific characteristics of news video structures and production techniques. It employs features like the appearance of anchor shot, long silence, blank frames and/or cue phrases etc to induce the story boundaries. The third class of methods employs machine learning techniques based on multi-modality features. Of course, any successful method needs to exploit both the statistical distribution of multimodality features and specific characteristics of news domain.

2.1 Text-based Techniques

Most groups [3, 10, 16] employed the text tiling method as the basis to perform story segmentation using the ASR output of news video. [16] employed a basic text-tiling method to perform the task, and could achieve an F_1 performance of only about 0.25. The run using a multi-resolution [3] variant of text tiling could achieve an F_1 performance of about 0.53. The run reported in [10] incorporates two additional features in an essentially text-tiling approach. First, they identified repeated nouns as topic words and link sentences containing them as belonging to same topic. Second, they extracted N-gram-based cue phrases that appear at the beginning of news stories. The results obtained with these additional features, however, have not been satisfactory. One reason is that the use of noun grouping is not reliable.

Another scheme [5, 8] uses a combination of decision tree and Maximum Entropy models. It takes a variety of lexical, semantic and structural features as inputs and generates boundary score at non-speech candidates, where no ASR words are found. It achieves an F_1 performance of 0.62.

2.2 Heuristic Rule-based Techniques

The use of purely text-based method is limited in accuracy because of two reasons. First, text alone is insufficient to determine the precise boundary of stories. Second, although deep linguistic discourse analysis offers great promise, the accurate location of discourse boundary is still an open problem. To overcome these problems, we must make better use of domain knowledge and other clues derived from audio-visual (AV) features. The essential domain knowledge that relates to news story boundary is the presence of anchor person shots, cue phrases, prosody, end of commercial and blank frames that often appear at the beginning of a new story.

One submission [15] used the simplest clues, the presence of blank frames plus specific detector for weather news and other heuristics rules, to induce story boundaries, and reported low accuracy. Another submission [4] used speech pause, cue phrases along with color histogram to extract shot boundaries. The algorithm declares a story boundary when both cue phrases and speech pause are found simultaneously in a shot and achieves 0.38 in F_1 measure.

Several groups centered their techniques based on the appearance of anchor person shots. Most employed a variety of shot clustering, face detection and likely positions of anchor persons to identify anchor person shots [12]. One scheme [14] used the transition from non-anchor person to anchor person shot, and from commercial to non-commercial shot as clues to story boundaries. They developed two separate classifiers based on heuristic rules and Maximum Entropy, and reported an F_1 measure of about 0.54.

2.3 Machine-Learning-based Techniques

In general, it is clear from Figure 1 that the use of combined AV and text features improves the performance of story boundary detection. This demonstrates the importance of incorporating additional AV features and domain knowledge in tackling this problem. Several entries have explored the use of fully automated machine learning techniques incorporating multimodality features. The entry of [2] employed features like anchor shot, face, motion activities, video text and other clues to identify story boundaries using a SVM method. They could achieve 0.38 in F_1 measure.

The approach described in [8] emphasizes a systematic solution to automatically learn the relevant feature subset and the optimal fusion model for each broadcast channel by using a flexible and efficient statistical framework, Maximum Entropy (ME). It uses a data-driven approach to avoid dependence on manual specification of domain-specific knowledge. The ME model [8] constructs an exponential log-linear function that fuses multiple binary features to approximate the posterior probability of an event (i.e., story boundary) given the audio, visual, or text data surrounding the candidate points. The binary features (195 in total) are produced using a feature wrapper function which takes multi-dimensional raw features as input and generates binary output. It takes into account different binarization thresholds, spatio-temporal sampling schemes, and diverse combinations of primitive features in different modalities.

In addition, a generic feature selection procedure is introduced to select the salient subset of features from the large feature pool. It employs an iterative greedy search process and measures the contribution of each feature by computing the K-L divergence between the empirical distribution and the likelihood function

based on the learned generative model. Starting from the original pool of 195 features, 35 most salient ones were automatically selected, with the top few being related to significant pause (combination of pause and pitch reset), anchor face, and speech segment transitions, etc. This is an interesting finding as the significant pause is a language- and gender- independent feature and is suitable for multi-lingual news segmentation. Earlier work in prosody analysis has also suggested its potential benefit [13]. The finding of other salient features like anchor face provides validation of some popular features used in earlier heuristics-based approaches. The above scheme achieves an F_1 performance of 0.69 with AV modalities and 0.74 with AV plus text modalities.

The scheme described in [3] uses a two-level, multi-modal framework to segment news video. They selected a combination of features include: visual-based features such as color; object-based features such as face, video-text; temporal features such as audio, motion and shot duration; and, text-based features such as the cue-phrases. They tackled the problem at two levels -- shot and story levels -- as illustrated in Figure 2. At the shot level, they classified the input shots into the semantic categories of *Anchor, 2Anchor, People, Speech/Interview, Sports, Finance, Weather,* etc by employing specific detectors for visual oriented categories and machine learning approach (Decision Tree) for the rest of categories. At the story segmentation level, they performed HMM (Hidden Markov Models) analysis to identify story boundaries by using features based on shot category information, scene change between shots and the presence of cue-phrases. They reported an F_1 performance of 0.75 using AV modalities and 0.77 using the combined AV plus text modalities. The system analyzes news video at the shot and then story levels, analogous to the idea from NLP that analyses text at the phrase and sentence levels. The use of two-level framework divides the problems into stages and partially alleviates the data sparseness problem in machine learning approaches.

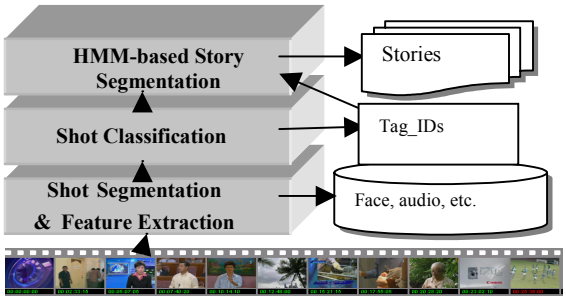

Figure 2: 2-level multi-modality framework [3]

3. OBSERVATIONS AND LESSONS LEARNT

3.1 Observations

The following observations can be derived from the results:

- The use of purely text-based method could achieve the best F_1 performance of around 0.62. The result is lower than what can be achieved by utilizing only the AV features, and the combined AV plus text features. The results dispel the general believe that good quality text alone is sufficient to achieve good performance in semantic task. In addition, the requirement of detecting story boundaries to within 5-second intervals further make the task more challenging for using the text evidence alone.
- The experience also suggests that the use of simple domain-motivated AV features like the presence of faces, anchor persons, blank frames, prosody etc, are effective to some extent. With the incorporation of text-based features such as the cue-phrases at the beginning of news stories, the performance of the system can generally be improved as evident from the top 2 performing systems.
- Machine learning techniques generally perform better than the heuristic rule-based techniques. The results suggest that with the judicious choice of features, the use of rigorous machine learning methods can be extremely effective over the use of hand-crafted heuristic rules. In fact, the current performance (F_1 measures 0.74 – 0.77) is approaching the usable level for practical systems. Models like HMM [3] prove to be promising for capturing dynamic structures in new video domains. Techniques such as ME feature induction [8] also offer a systematic avenue for automatically discovering the effective features (low-level as well as mid-level). Such automatic capabilities are important for extending the results to new data sources such as foreign news.

3.2 Lessons Learnt

There are two main categories of errors. The first is in feature extraction while the second is in story segmentation. For feature extraction, our analysis showed that *face, motion* and *audio classes* are some of the most important features, and *audio* accounts for most of the feature extraction errors. The errors in low-level feature extraction affect the accuracy of extracting mid-level features like the *anchor-person shot*, which is perhaps the most important mid-level feature in story boundary detection task. Some groups [3, 14] could achieve an *anchor-person* detection accuracy of between 60-80%. Our recent experiment reveals that if the detection accuracy of *anchor-person* shot is closed to 100%, we can achieve an F_1 measure of about 0.62 by using only this feature.

For story segmentation, there are at least two sources of errors. First, we found that there are many story patterns that were not discovered by the machine learning methods during training. One possible reason is that such "unexpected" patterns did not occur sufficiently frequently in training data for ML methods to learn the patterns, which is a case of data sparseness problem. For example, 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 as a story unit. 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 5% of errors in story segmentation. Another possible reason is that the features adopted in the detection system may not be sufficient to support some patterns. The second source of error is related to the segmentation of news stories into sub-stories, such as the segmentation of sports news into individual sport news (football, basketball, etc.). In [9], authors compare fusion capabilities of generative, discriminative, and ensemble learning models on multi-modal feature sets and carefully analyzed the error cases according to different video genres. Among them, two story types – sports and news briefing without visual anchor segments stand out as they suffer from a

significant miss rate (over 50%) and jointly account for 36% of story boundaries in the CNN videos. Such cases remain to be challenging as the data usually has short durations and/or rapid temporal.

One approach to overcome the above two problems is to add control mechanisms in the machine learning framework to incorporate domain knowledge instead of relying on a purely data driven approach. An alternative solution is to investigate the use of higher order statistics that could learn complex patterns and handle hierarchical categorization. Some recent progresses in Hierarchical HMM and genre-adaptive dynamic models offer promising directions for further investigation. Finally other high-level features such as Video OCR may be used to expand the power of the feature pool.

4. CONCLUSIONS

This paper reviewed the recent efforts in TRECVID 2003 on segmenting over 120-hour of news video into story units. We categorized the technical approaches used by different groups and analyze the types and sources of errors. The general conclusion is that it is advantageous to employ rigorous machine learning techniques, along with judicious use of full multi-modality features, to achieve good segmentation performance. In particular, systematic methods for feature selection and temporal dynamic modeling are rewarding. The results also point to several directions for future work. One direction is to explore the use of higher order statistical techniques such as the hierarchical HMM or multi-level classification framework [3]. Another direction is to explore the genre-adaptive dynamic system to model the structure variation in different genres of stories

5. References

1. J. Allan, J. Carbonell, G. Doddington, J. Yamron & Y. Yang (1998). Topic detection and tracking pilot study final report. Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. 194-218.
2. P. Browne, C. Czirjek, G. Gaughan, C. Gurrin, G. J.F. Jones, H. Lee, S. Marlow, K. Mac Donald, N. Murphy, N. E. O'Connor, N. Hare, A. F. Smeaton, & J. Ye (2003). Dublin City University video track experiments for TREC 2003. Notebook submitted to TRECVID 2003.
3. L. Chaisorn, T.-S Chua, C.-K Koh, Y.-L Zhao, H. Xu, H. Feng & Q. Tian (2003). A two-level multi-modal approach for story segmentation of large news video corpus, Proceedings of TRECVID workshop 2003.
4. D. Eichmann & D.-J. Park (2003). Experiments in boundaries recognition at the University of Iowa. Notebook submitted to TRECVID 2003.
5. M. Franz, J. S. McCarley, S. Roukos, T. Ward, and W.-J. Zhu,"Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering broadcast news domain," in Proceedings of TDT-3 Workshop, 2000.
6. J.L. Gauvain, L. Lamel & G. Adda (2002). The LIMSI broadcast news transcription system. Speech Communication, 37(1-2):89-108.
7. M.A. Hearst (1994). Multi-paragraph segmentation of expository text. Proc of the 32nd Annual Meeting of the Association for Computational Linguistics.
8. W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in IS&T/SPIE Electronic Imaging, San Jose, CA, 2004.
9. W. Hsu and S.-F. Chang, "Generative, Discriminative, and Ensemble Learning on Multi-modal Perceptual Fusion toward News Video Story Segmentation," IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, June 27-30, 2004.
10. P. Rennert (2003). StreameSage unsupervised ASR-based topic segmentation. Proceedings of TRECVID workshop.
11. J.C. Reynar (1994). An automatic method of finding topic boundaries. Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics.
12. TREC Video Retrieval Evaluation (2003), http://www-nlpir.nist.gov/projects/tv2003/tv2003.html. Washington D.C., 17-18 Nov, 2003.
13. J. Vaissiere, "Language-independent prosodic features," in Prosody: Models and Measurements, Anne Cutler and D. Robert Ladd, Eds., pp. 53-66. Springer, Berlin, 1983.
14. L. Wu, Y. Guo, X. Qiu, Z. Feng, J. Rong, W. Jin, D. Zhou, R. Wang & M. Jing (2003). Fudan University at TRECVID 2003. Notebook submitted to TRECVID 2003.
15. Y. Zhai, Z. Rasheed & M. Shah (2003). Univerisity of Central Florida at TRECVID 2003. Notebook submitted to TRECVID 2003.
16. .Zugano, K. Hoashi, K. Mutsumato, F. Sugaya & Y. Nakajima (2003). Shot boundary determination on MPEG compressed domain and story segmentation experiments for TRECVID 2003. Notebook in TRECVID 2003.