A DETECTION-BASED APPROACH TO BROADCAST NEWS VIDEO STORY SEGMENTATION

Chengyuan Ma, Byungki Byun, Ilseo Kim and Chin-Hui Lee

School of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332, USA
{cyma, yorke3, ilseo, chl}@ece.gatech.edu

ABSTRACT

A detection-based paradigm decomposes a complex system into small pieces, solves each subproblem one by one, and combines the collected evidence to obtain a final solution. In this study of video story segmentation, a set of key events are first detected from heterogeneous multimedia signal sources, including a large scale concept ontology for images, text generated from automatic speech recognition systems, features extracted from audio track, and high-level video transcriptions. Then a discriminative evidence fusion scheme is investigated. We use the maximum figure-of-merit learning approach to directly optimize the performance metrics used in system evaluation, such as precision, recall, and F1 measure. Some experimental evaluations conducted on the TRECVID 2003 dataset demonstrate the effectiveness of the proposed detection-based paradigm. The proposed framework facilitates flexible combination and extensions of event detector design and evidence fusion to enable other related video applications.

Index Terms— event detection, discriminative fusion

1. INTRODUCTION

Broadcast news video is well structured, from frames to shots, scenes, and stories. The story boundaries are important information, because the story is a basic unit of many multimedia indexing, retrieval, and management systems. Numerous studies on video story segmentation have been conducted using different knowledge representations and machine learning algorithms based on various design strategies.

The *Informedia* system [1] was one of the early rule-based systems. Some ad hoc rules were designed to combine visual, acoustic, and textual features. Other state-of-the-art video story segmentation systems are based on statistical modeling approaches. Within the Shannon's channel decoding framework for pattern recognition, some algorithms have been investigated. For instance, the broadcast news was modeled with a hidden Markov model (HMM) [2] [3] or a probabilistic context-free grammar (PCFG) [4] such that the story boundaries and other complicated structure information can be obtained by a decoding procedure or a parsing tree. For a HMM or PCFG based system, tokenization of the multimedia stream is crucial. The story segmentation performance greatly depends on the selection of the tokens to represent the video structures. For example, 17 predefined shot categories are used to capture the structure information [2]. Some of them are program specific logos, such as "SPORT" and "TOP". These tokens are helpful in finding the structures of broadcast news videos. However, its limitation is obvious: the production rules and the style of the broadcast news video vary over programs and time. So a complete and accurate channel specification that is the key to the success of Shannon's channel decoding paradigm cannot be easily realized for such a complex system, which deals with so many diverse sources of information that cannot be handled in an integrated manner.

Inspired by studies of human visual [5] and auditory perception [6], a detection-based framework was proposed to handle the environment variabilities and diverse knowledge sources in automatic speech recognition [7]. By decomposing a big problem into small pieces, a divide-and-conquer strategy provides an attractive and feasible solution. Each subproblem can be solved with different design instead of a single feature or a single likelihood computation. By solving smaller pieces one by one, the evidence collected from subproblems can be combined to obtain a solution of a complex system. So for video story segmentation with many diverse and heterogeneous information sources, a detection-based framework provides a natural and intuitive solution. In a detection-based framework, story segmentation is divided into two steps: event detection and evidence fusion. First, a collection of possible story boundaries were detected as candidates. And then the evidence around each candidate point was collected and some fusion approaches were used to combine the heterogeneous information. For instance, support vector machine (SVM) [8] and maximum entropy (MaxEnt) model [9] have been successfully investigated.

Based on the statistical detection and decision theory, different signal processing and machine learning algorithms are employed to detect multi-modal events. As for evidence fusion, the simplest way is to concatenate the low-level features from different sources into a large vector. However, from previous studies, it is well known that some features are more informative in story segmentation, such as anchor shots, than other pieces of video features. These features should be treated differently. In this situation, fusion at the score level can be a better choice.

Another problem of the existing story segmentation systems is that the optimization criteria in the training phase is inconsistent with the performance metric used in the evaluation phase. This problem becomes more severe when dealing with imbalanced data, where negative instances will generally dominate both the training and evaluation data sets. For example, in video story segmentation, the portion of the true story boundaries in all the candidate story boundaries is about 4%. The widely used up-sampling with replacement of positive instances or down-sampling of negative samples will bring some difficulties. When re-sampling is finished, the two group of instances tend to have equal number of examples or have a pre-fixed ratio. However, the class prior information was lost.

In this paper, we present a detection-based video story segmentation system. It includes multi-modal event detection and a discriminative evidence fusion scheme. The maximum figure-of-merit (MFoM) learning approach [10] is used for evidence combination, which directly optimizes the performance metrics used in video story segmentation, such as precision, recall, and F1 measure. Another benefit for using MFoM learning is that in previous studies, this approach demonstrated good performance for imbalanced data [10]. Our experimental result on TRECVID 2003 dataset also showed its effectiveness for video story segmentation.

2. KNOWLEDGE SOURCES AND EVENT DETECTION

The shot is a basic unit of a story segmentation system and the shot boundaries are used as candidate points. Our preliminary experimental results show that 93.3% of the true story boundaries are covered by the shot boundaries. From the context of each candidate boundary, there is rich of information from heterogeneous knowledge sources. Many different multi-modal event detectors can be constructed. Detector design is similar to the matched filter design for artificial signals in communication systems. It is more difficult to design detector for natural signals like speech and video. Statistical detection and decision theory show that under different conditions, we can design detectors using different criteria, e.g., Bayes criterion, Min-Max criterion, and Neyman-Pearson criterion [11]. And the commonly used probability of error criterion is a special case of Bayes criterion. Next we describe four evidence detectors using the visual, acoustic, and textual features respectively.

2.1. Anchor shot detection

Our previous studies on a 17-hour broadcast new video dataset showed that if an anchor shot location is treated as a story boundary, this feature alone could achieve an accuracy of 61.7% in story segmentation [12]. Similar results have been verified by many other studies [13]. Research work on feature selection for story segmentation using an information gain criterion also shows that anchor shot is the most important feature in story segmentation [3]. Our previous study demonstrated an unsupervised anchor shot detection method almost approaches the supervised anchor shot detection performance. In this paper, to get the best story segmentation performance, a supervised anchor shot detection system was constructed as described in [12]. A probabilistic confidence is obtained for each shot instead of a hard decision.

2.2. LSI-based AIA detectors

Semantic concept detectors can help to bridge the gap between the low level features and the high-level semantics. Previous studies showed that the semantic concepts provide important information of story boundaries [2]. In addition, the dynamic patterns of semantic concepts showed the structure of a broadcast news video. Large-scale concept ontology for multimedia (LSCOM) [14] is designed to provide a dictionary of semantic units for general purpose indexing and retrieval of video. These concepts were used for story segmentation in this paper.

First, 43 semantic concepts were manually selected and these concepts are shown to be closely related to the story boundaries. For each candidate point, a change of the detected concepts often indicates a potential story boundary. A latent semantic indexing (LSI)-based automatic image annotation (AIA) system was constructed as described in [15]. This system was verified to perform very well in the AIA task for the Corel photo dataset. As the training and validation set, we used the TRECVID 2005 development set for which all the LSCOM semantic concepts were annotated [14]. Every shot in the TRECVID 2003 used for the story boundary segmentation task was then annotated with the AIA-based detectors. One thing to note is that the video data in TRECVID 2003 and in TRECVID 2005 have large mismatches in video qualities. So instead of creating a hard-decision for each shot, a confidence measure is given by each concept detector. The confidence vectors are combined with the proposed discriminative fusion method.

2.3. Audio type detection

There are many types of audio signals in the audio track of broadcast news video, e.g, speech, music, silence, and speech with music background, etc. Both the type of audio signal and the change of the audio type imply important information for story segmentation. In our study, each audio type is modeled with a HMM and each shot is represented by a confidence vector of audio type. This is similar to the confidence vectors in AIA detectors.

2.4. Cue-phrase detection

Intuitively, there are some cue phrases around the story transition. The automatic speech recognition (ASR) transcriptions were used to detect the cue phrases within each shot period. To extract a list of cue phrases, we compute the N-gram lexical sequences around reference story boundaries. From this list, 16 transition word sequences are manually selected, e.g., "ABC news", "CNN", "thanks for watching", "coming up", and "ahead on", etc. To be suited in the evidence fusion framework, the detected cue phrases are converted into a score according to the occurrence frequencies. With more cue phrases detected within a shot, the confidence is higher.

3. DISCRIMINATIVE EVIDENCE FUSION

When the evidence from diverse knowledge sources around each candidate story boundary is available, some fusion strategies could be employed to obtain a final solution by integrating heterogeneous sources. Unlike other fusion approaches such as maximum entropy (MaxEnt) method that maximizes the likelihood, and SVM that maximize the margin between decision boundaries, the maximum figure-of-merit (MFoM) learning algorithm directly optimize the performance metrics used in video story segmentation evaluation: precision, recall, and F1 measure.

3.1. Maximum figure-of-merit (MFoM) learning

The maximum figure-of-merit learning approach [10] [16] approximates the four terms in the contingency table with a differentiable function with regards to the model parameters and directly optimizes the performance metrics. It allows quite a bit of flexibility in choosing the discriminant functions for each class. The discriminant functions $f(\mathbf{x};\mathbf{w})$ can be linear discriminant functions (LDF), quadratic discriminant functions (QDF), or complicated probabilistic discriminant functions such as Gaussian mixture model (GMM) and HMM. With the help of the discriminant functions for both positive (boundary) and negative classes (non-boundary), a misclassification measure $d(\mathbf{x};\mathbf{w})$ can be used to measure the correctness of a decision at a single candidate point. Here $\mathbf{x}$ and $\mathbf{w}$ are evidence vector and parameters of the discriminant functions respectively. $\mathrm{X}_{\mathrm{pos}}$ and $\mathrm{X}_{\mathrm{neg}}$

are training instances from positive and negative classes.

$$d(\mathbf{x};\mathbf{w}) \quad = \quad \begin{cases} f_{\text{neg}}(\mathbf{x};\mathbf{w}_{\text{neg}}) - f_{\text{pos}}(\mathbf{x};\mathbf{w}_{\text{pos}}), & \mathbf{x} \in \mathrm{X}_{\text{pos}} \\ f_{\text{pos}}(\mathbf{x};\mathbf{w}_{\text{pos}}) - f_{\text{neg}}(\mathbf{x};\mathbf{w}_{\text{neg}}), & \mathbf{x} \in \mathrm{X}_{\text{neg}} \end{cases} \quad (1)$$

The role of a loss function $l(\mathbf{x};\mathbf{w})$ is to use a smooth function to approximate the errors obtained on a dataset. And the sigmoid function is the most widely used one. Here α control the steepness of the curve and β control the covered area of the loss function.

$$l(\mathbf{x};\mathbf{w}) \quad = \quad \frac{1}{1+\exp\left(\beta - \alpha * d(\mathbf{x};\mathbf{w})\right)} \quad (2)$$

The performance metrics like error rate, precision, recall, and F1 measure can be calculated from the contingency table as shown in Table 1. Here, true positive (TP), false positive (FP), false negative (FN), and true negative (TN) are approximated by the loss function for each training sample.

Table 1. Contingency table.

	test (+)	test (-)
+	TP $\approx \sum_{\mathbf{x}\in \mathrm{X}_{\text{pos}}} (1 - l(\mathbf{x};\mathbf{w}))$	FN $\approx \sum_{\mathbf{x}\in \mathrm{X}_{\text{pos}}} l(\mathbf{x};\mathbf{w})$
-	FP $\approx \sum_{\mathbf{x}\in \mathrm{X}_{\text{neg}}} l(\mathbf{x};\mathbf{w})$	TN $\approx \sum_{\mathbf{x}\in \mathrm{X}_{\text{neg}}} (1 - l(\mathbf{x};\mathbf{w}))$

Given the approximation of the four terms in a contingency table, the F1 measure can be approximated as follow.

$$\text{F1} \quad = \quad \frac{2\text{TP}}{\text{FP} + \text{FN} + 2\text{TP}} \quad (3)$$

3.2. Parameter initialization and update

An issue needs to be considered is the initialization of the optimization procedure. Because generally, the approximated F1 measure is a non-convex function with regards to the model parameters, there is no guarantee to find a global optimal solution. And the obtained local optimal solution greatly depends on the initial value of the parameters. The initialization in our experiments is performed using the expectation maximization (EM) algorithm [17] for GMM and perceptron algorithm [18] for LDF. The parameter update can be conducted using batch gradient descent method, or generalized probabilistic descent method, or quasi-Newton optimization techniques (e.g., L-BFGS).

3.3. Discriminative evidence fusion

For each candidate point, there are four evidence scores from different detectors. For example, if the shot right after the candidate point is detected as an anchor shot with a high confidence and there are some cue phrases detected right before the candidate point, this candidate point is very likely to be a true story boundary. When some evidence is detected with a hard-decision, some logical rules can be deduced and constructed for video story segmentation. In this paper, GMM discriminant functions are used in AIA feature fusion and LDFs are used in story segmentation fusion. The MFoM learning approach is used to improve the performance of the final decision. This data-driven approach will learn a weight for each evidence and an offset from both positive the negative instances. It allows quite a bit flexibility in evidence detector design. When some new detectors are built, their relative importance for final decision making will be learned automatically and discriminatively from data. The inference step is really straightforward.

4. EXPERIMENT SETUP AND RESULT ANALYSIS

All experiments were conducted on a standard benchmark dataset. The complete dataset of TRECVID 2003 [19] is used for story segmentation evaluation. It consists of about 110 video clips for development and another 105 video programs for evaluation. Each video clip has a length of about 30 minutes. The data are CNN and ABC broadcast news video of year 1998. We have conducted our experiments using the methodology proposed by TRECVID [19].

The shot segmentation, keyframe extraction, and ground-truth story segmentation were provided by LDC [20]. The audio track was demultiplexed from the MPEG stream with 16 KHz sampling rate and 16 bits. Mel frequency cepstral coefficient (MFCC) features were extracted from audio signals in the audio type detector design [21].

The performance of a story segmentation system is usually evaluated with precision and recall. Meanwhile, a single F1 measure, which is the harmonic mean of the recall and precision, is also used for performance comparison. According to the guideline of TRECVID 2003, when conducting performance evaluation, each reference boundary was expanded with a fuzziness factor of 5 seconds in each direction, resulting an evaluation interval of 10 seconds.

4.1. Comparison with SVM fusion

The first experiment is to compare the performance of the MFoM fusion method with the SVM fusion method using the AIA featuers. For each candidate point, a confidence vector of 86 dimension is constructed using the AIA detector output for the shot right before and after the candidate boundary. Two fusion strategies have been investigated. The first one is a SVM fusion approach using the LIBSVM [22] toolkit. And the second one is the MFoM scheme. In the MFoM scheme, both the positive class (boundary) and negative class (non-boundary) are modeled by a GMM with 2 mixtures. Figure 1 showed the result of story segmentation using this AIA confidence vector. It's clear that MFoM significantly outperformed the SVM fusion method. The F1 measure from the MFoM scheme is about 0.51 and the F1 measure from SVM is about 0.44.

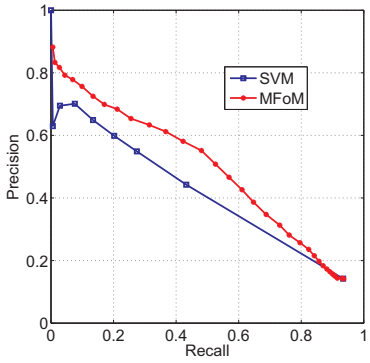

Fig. 1. Comparison of SVM fusion and MFoM fusion.

4.2. Heterogeneous information source fusion

The second experiment is to demonstrate the effectiveness of MFoM fusion scheme with more heterogeneous information sources. Table 2 shows the performance of story segmentation under different combinations. The upper part of the table shows the performance using individual detectors. For instance, using only AIA detectors can achieve a F1 measure about 0.514, while using only anchor detector can achieve a F1 measure about 0.605. It's also clear that anchor detector has a high precision for story segmentation. It means that the evidence from the anchor detector is a reliable cue. Nevertheless, audio type detectors demonstrated a high recall. It means that audio type changes can cover most of the true story boundaries. These kinds of complementary information gives the detection-based story segmentation system plenty of room for performance improvement. In this detection-based framework, when more and more evidences are available, the system performance can be improved in an additive manner in terms of F1 measure. For example, when the cue phrase detector was combined with the anchor detector, even a simple "OR" operation can improve the F1 measure by 1% from a high performance baseline system. Similar experiments have been done with AIA detectors and audio type detectors in the bottom part of Table 2. It's obvious that with more evidences were combined using MFoM scheme, the recall was greatly increased while the precision was decreased a little.

Table 2. Performance of story segmentation.

	Precision	Recall	F1
Text (T)	0.382	0.208	0.269
Audio (A)	0.194	0.771	0.310
AIA (V)	0.552	0.481	0.514
anchor	0.780	0.494	0.605
anchor + T	0.762	0.520	0.618
anchor + T + V	0.753	0.552	0.637
anchor + T + V + A	0.739	0.581	0.651

5. SUMMARY

In this paper, we present a detection-based framework for pattern recognition and demonstrate its application to video broadcast news story segmentation. In the proposed approach, event detectors and fusion schemes can be designed and optimized separately. Therefore, it offers quite a bit of flexibility in system construction and performance improvement. With more and more event detectors available, system performance can be improved in an additive manner. This study also presents a set of multi-modal event detectors built with different signal processing and machine learning methods, and a discriminative fusion method for video story segmentation, which directly optimizes F1 measure. There are some useful event detectors not covered in this paper. For future work, we will add more candidate points to the boundary list such that the maximal recall from the candidate set can approach 100%. In addition, speaker change detection, visual and acoustic scene change detectors are critical for further performance improvement.

6. REFERENCES

[1] A. Hauptmann and M. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Proc. of Advances in Digital Libraries*, 1998.

[2] L. Chaisorn, T.-S. Chua, C.-H. Lee, and Q. Tian, "A hierarchical approach to story segmentation of large broadcast new video corpus," in *Proc. of ICME*, 2004.

[3] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proc. of ACMMM*, 2004.

[4] A. Jacobs, G. T. Ioannidis, S. Christodoulakis, N. Moumoutzis, S. Georgoulakis, and Y. Papachristoudis, "Automatic, context-of-capture-based categorization, structure detection and segmentation of news telecasts," in *DELOS Conference*, 2007.

[5] S. W. Zucker, "Computer vision and human perception: An essay on the discovery of constraints," in *Proc. of IJCAI*, 1981.

[6] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. SAP*, vol. 2, pp. 567–577, 1994.

[7] C. Ma, Y. Tsao, and C.-H. Lee, "A study on detection based automatic speech recognition," in *Proc. of InterSpeech*, 2006.

[8] W. Hsu, L. S. Kennedy, S.-F. Chang, M. Franz, and J. R. Smith, "Columbia-IBM news video story segmentation in TRECVID 2004," Tech. Rep., Columbia University, 2005.

[9] W. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *Proc. of ICME*, 2003.

[10] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A MFoM learning approach to robust multiclass multi-label text categorization," in *Proc. of ICML*, 2004.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing. Volume II: Detection Theory*, Prentice Hall, 1998.

[12] C. Ma and C.-H. Lee, "Unsupervised anchor shot detection using multi-modal spectral clustering," in *Proc. of ICASSP*, 2008.

[13] A. Yanagawa, W. Hsu, and S.-F. Chang, "Anchor shot detection in TRECVID-2005 broadcast news videos," Tech. Rep., Columbia University, 2005.

[14] L. Kennedy and A. Hauptmann, "LSCOM lexicon definition and annotation version 1.0," Tech. Rep., Columbia University, 2006.

[15] B. Byun, C. Ma, and C.-H. Lee, "An experimental study on discriminative concept classifier combination for TRECVID high-level feature extraction," in *Proc. of ICIP*, 2008.

[16] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization," *ACM Trans. Inf. Syst.*, vol. 24, no. 2, pp. 190–218, 2006.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000.

[19] A. F. Smeaton, W. Kraaij, and P. Over, "TRECVID 2003 - an overview," Tech. Rep., National Institute of Standards and Technology, 2004.

[20] G. Quenot, C. Petersohn, P. Over, and K. Walker, *TRECVID 2003 Keyframes & Transcripts*, LDC, 2007.

[21] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[22] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.