BROADCAST NEWS STORY BOUNDARY DETECTION USING VISUAL, AUDIO AND TEXT FEATURES

Maryam Daneshi, Matt Yu

ABSTRACT

News video story segmentation is vital for video summarization, story linking, and curation. We present a multimodal segmentation algorithm which fuses video, audio and text cues for story boundary detection. We show that broadcast news closed captioning is a rich and readily available source that improves story boundary detection. Furthermore, we propose an empirical distribution-based feature representation for various binary video and text features. We investigate different multimodal fusion methods for learning algorithms based on discriminative models and evaluate the performance of each set of features on more than 55 hours of three major news programs. We compare the effect of each feature and fusion method on the segmentation accuracy of each news program.

Index Terms— story segmentation, multimodal fusion, supervised learning

1 Introduction

News video story segmentation, a process which breaks news streams into individual stories, is an important tool for next generation news systems. However, the wide variation of production rules across channels and diversity of stories in each program make segmentation a difficult task.

To overcome this difficulty, it is essential to take advantage of the different modalities in news streams: video, audio, and closed captioning. However, many useful cues are inherently binary, e.g. the existence of an anchor in a shot. Furthermore, gathering training data can be problematic because shots which do not contain story boundaries severely outnumber shots that do contain story boundaries.

In this paper, we describe a novel story segmentation algorithm which extracts and fuses features from multiple cues and trains a story boundary classifier. Given training data for a news program, we generate a separate classifier which learns the story boundary characteristics of that program. To adapt our binary features, we propose an empirical distribution-based feature generation model which builds a conditional distribution for the distance of each feature from a story boundary. While these cues, individually, are insufficient to reliably indicate a story boundary, an intelligent combination of these cues can produce a strong indicator. To that end, we also study different fusion methods in our discriminative classification approach and analyze the performance of each technique. Finally, to deal with the imbalance between non-story and story sample points in the training data, we use a bootstrapping approach.

The rest of the paper is organized as follows. Sec. 2 reviews existing work on story boundary detection. In Section 3 we introduce all visual, audio and text cues as well as the classification framework. Finally in Sec. 4 we present our experimental results and compare the performance of various features and fusion methods.

2 Related Work

In previous work, researchers detected story boundaries by combining multiple cues. However, the majority of these efforts focused on just video and audio features. Hsu et al. [1] considered the appearance of an anchor person, audio pitch jump and significant audio pauses as the main set of features. Zhai et al. [2] obtained text from Automatic Speech Recognition (ASR) to detect potential segmentation points.

Multimodal feature fusion has been of great interest to the research community. Hsu et al. [3] used a maximum entropy objective to select the most informative mid-level audio and video features and demonstrated an optimal feature fusion method. In later works, Hsu et al. [4, 5] investigated alternative discriminative models, i.e. Support Vector Machine (SVM), and showed a performance improvement when combining maximum entropy with SVM. While Jianping et al. [6] presented a Naive Bayes approach for story boundary detection, Gao et al. [7] combined syntactic and semantic methods for segmentation using an unsupervised learning method.

3 News Story Segmentation

Figure 1 shows an overview of the features we use for segmentation. In Sec. 3.1 we describe the video and audio features followed by a detailed description of the text features in Sec. 3.2. Later, in Sec. 3.3, we describe our multimodal story segmentation algorithm.

3.1 Video and Audio Features

3.1.1 Anchor Frames

Usually, story transitions occur while an anchor is visible or has recently appeared. Furthermore, anchors appear more often than any other person in the news and move very little. Thus, to detect anchor shots, the following method is used: news video is sampled every $200\ ms$ and both face detec-

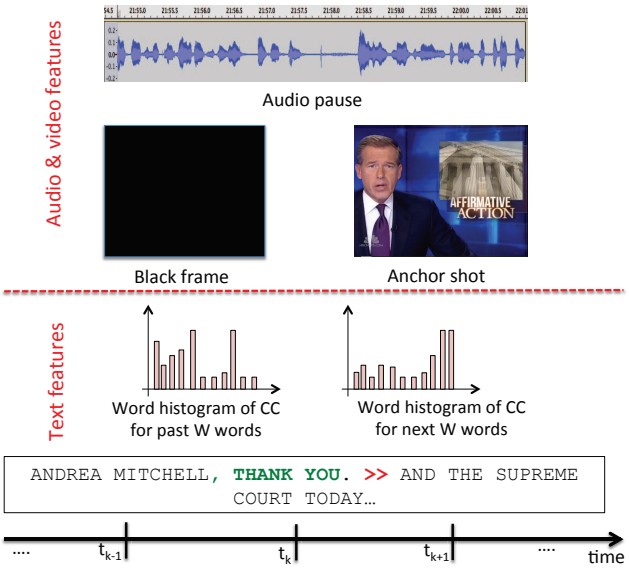

Fig. 1: Segmentation features overview - for any time in the news program various video, audio and text features are collected

tion and tracking are used to find all face appearances. In the detection phase, a Viola-Jones frontal-face detector [8] is applied to find all faces in a frame. Then a color histogram based tracker, Camshift [9], is applied to track all detected faces in the following frames. In the last stage, we cluster all the color histograms using L1 distance. The cluster with the shots that are closest to the dominant cluster centroids are the anchor shots.

Furthermore, most news programs tend to follow the same pattern during a story transition, i.e the anchor appears on the left with an image of the next story over his shoulder. Thus, the geometric location of the bounding box for the detected face is also used as a visual cue.

3.1.2 Black Frames

It has been shown in [10] that commercials are usually followed by one or two frames that are entirely black. For broadcast news, we observe that a new story starts after a commercial. However, a black frame may also appear in other parts of the news [11]. Therefore, like the other cues, black frames are not, by themselves, a reliable indicator of a segment boundary. Instead, they provide added information to improve the segmentaiton.

3.1.3 Significant Pause Detection

Long pauses were shown in [5] to be good indicators of story boundaries. This audio cue captures the tendency of an anchorperson to pause momentarily before introducing a new story. At every time point of interest, t, the longest period of silence, $p(t)$, is calculated in the interval $[t-W, t]$, where W is the sampling period. A period of silence is defined as a time duration where the audio volume, $V(t)$, is always below half the average volume of the news stream, τ. Mathematically:

$$p(t) = \max_{\substack{t_2 \in [t-W,t], \\ t_1 < t_2, \\ V(t_k) < \tau,\ \forall t_k \in [t_1, t_2]}} (t_2 - t_1)$$

3.2 Text Features

By U.S. law, closed captioning (CC) must be provided with all news videos. Often, >>> and >> markers are inserted to denote story changes or speaker changes, respectively. Since CC text is transcribed by a human, it is always delayed relative to the video. To obtain accurate timestamps for the CC words, ASR is performed on the audio track using the CMU Sphinx Toolkit [12]. Words in CC and ASR are matched and aligned using a dynamic time warping algorithm [13, 14], yielding timestamps which have an average error of less than a second. ASR recognized words are then matched with CC words to align the entire CC text.

3.2.1 Bag of Words Histogram Distance

Consecutive stories tend to contain many different words. At any time point, we calculate two bag-of-words histograms with tf-idf normalization, one for the words before the time point and one for the words after the sampled time. Then, we generate a continuous feature by calculating the distance between the two histograms:

$$f(t) = ||\mathbf{H}(t-W, t) - \mathbf{H}(t, t+W)||_1$$

where $f(t)$ is the distance between histograms of length W words at time t and $\mathbf{H}(t_1, t_2)$ is the bag-of-words histogram between times t_1 and t_2. To pick a meaningful window size W, the normalized histogram distance of the story boundaries are analysed on three news programs. From our analysis, a W value of 400 was the best window size.

3.2.2 Transition Phrase

Broadcast news writers use transitions to carry the audience from one story to another. We collected and built a database of over 600 reliable transition phrases from several months of news programs. At any time point, we collect a binary feature indicating the presence of any of these transitions phrases.

3.2.3 Reporter Change Marker

>>> and >> markers are inserted to denote changes in stories or changes in speakers, respectively. However, some news broadcasters use >> for both speaker and topic changes and not the story change marker at all, e. g. PBS News Hour. We define a binary feature that indicates the presence of >>.

3.3 Segmentation Framework

Many of our features are binary. Moreover, the presence of each of these features, does not directly imply a story boundary. For example, a story boundary may appear a few seconds after the appearance of a black frame. We define $\theta \in \{0, 1\}$ a

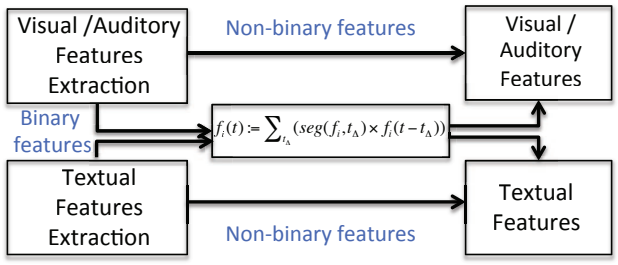

Fig. 2: Feature extraction-Binary features are enhanced using a prior empirical-based probability distribution

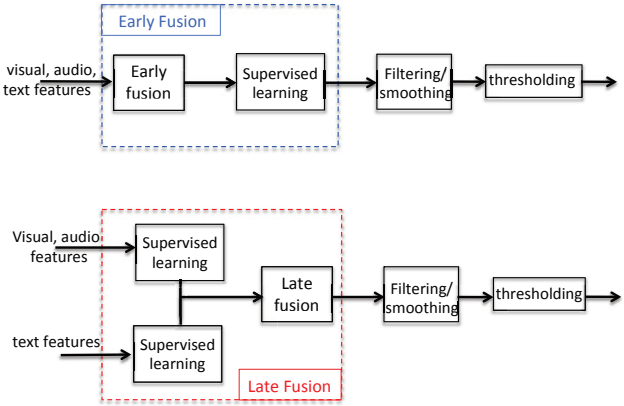

Fig. 3: Classification framework- Early (top) and late (bottom) fusion techniques are used in the evaluation framework followed by smoothing and filtering

random variable indicating a story boundary, and then calculate an empirical prior probability distribution of, $seg(f_i, t_\Delta)$, for all binary features as:

$$seg(f_i, t_\Delta) = p(\theta = 1 | f_i, t_\Delta) = \frac{\sum_t \left(g(t) \cdot f_i(t - t_\Delta)\right)}{\sum_t g(t) \cdot \sum_t f_i(t)}$$

where $f_i : T \rightarrow \{0, 1\}$ is the binary feature value at a given time and i is the feature index. t_Δ (in seconds) denotes the time difference between the feature indicator and the actual story boundary. $g : T \rightarrow \{0, 1\}$ represents the ground truth story segmentation data.

As shown in Fig. 2, first, the features described in Sec. 3.1 and Sec. 3.2 are extracted from the news video stream. Then, binary features are enhanced by the previously calculated empirical prior probabilities, seg, as:

$$f_i(t) := \sum_{t_\Delta} seg(f_i, t_\Delta) \times f_i(t - t_\Delta)$$

Figure 3 shows the evaluation approaches in our classification framework. We combine features into multimodal representations using both early and late fusion schemes [15]. In early fusion, the normalized feature vectors are concatenated and a Support Vector Machine (SVM) is used to classify time points as boundaries or non-boundaries. We use the Radial Basis Function (RBF) as the kernel function for the SVM:

$$K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2}$$

where $\gamma > 0, \gamma \in \Re$ is the kernel parameter. The γ parameter is set to be the inverse of the median of all the pairwise training instance distances [16]. Five fold cross validation is applied by varying C, the soft margin parameter of the SVM, on the training set to find the parameter which achieves the highest accuracy.

In late fusion, the same SVM is applied to the concatenation of video-audio features and the text features separately. The decision values (DV_{AV} and DV_T) of the classifiers are normalized and a weighted sum is calculated as:

$$DV_{LF} = \omega \times DV_{AV} + (1 - \omega) \times DV_T$$

$$0 \leq \omega \leq 1$$

where ω is found by exhaustive search to maximize the area under the Precision-Recall curve (average precision value). In this work, Precision and Recall are defined as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$ respectively. In our training and testing data, news videos are sampled at one second intervals, which leads to a highly imbalanced dataset. For example, in a 30 minute program with 7 stories, there are 7 positive samples and 1793 negative samples. Therefore, less than 1% of the labeled points are positive in this program. To overcome this imbalance, "hard negative" examples in the training set are isolated by selecting 20% of the negative and all the positive training samples to train a classifier. Then we test the classifier on the remaining 80% of the negative samples. Decision values of each of the tested samples are stored. After five iterations of the method above, the top 10% of the hard negative examples (samples with high false positive decision values) and all positive examples are picked as training samples. This method reduces the ratio of the positive to negative training samples to $1 : 10$. For our discriminative model we use manually annotated reference story boundaries. However, features are usually asynchronous across modalities and human annotated ground truth data is not very precise. Thus, in our data, sample points within a 3.5-second fuzzy window of the human annotations are also labeled as positive points.

Also, as a post classification step, we apply smoothing and filtering on the classifier's decision values. The smoothing is done using a lowpass filtered with a window size of 7 and the filtering uses a one-dimensional Laplacian filter with a kernel of:

$$[-1, -1, 2, 2, -1, -1]$$

The length of the smoothing and filtering kernels are chosen to be the same as the length of the ground-truth fuzzy window size we added to our sample point. The filtering and

smoothing steps highlight the regions of decision values with rapid change in a smoothed window around the potential story boundary points. This approach can improve the accuracy of the detection algorithm up to 5%. Finally, video sample points are classified to story and non-story by thresholding the decision values.

4 Experimental Results

We built a database of news streams from three major U.S. broadcast news programs. The database consists of 20 hours of NBC Nightly News, 24 hours of ABC world News and 11 hours of PBS News Hour collected from September to mid November of 2012. We used human annotators to create story boundary ground truths. On average there are seven to eight stories in every news program. We use 80% of the videos of each channel for training and the rest for testing. We sampled each news stream at one second intervals and extracted the video, audio and text features mentioned in Sec. 3

Table 1*:* Story boundary detection performance of each feature for 3 different news programs

	NBC News			ABC News			PBS News		
Feature	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
anchor(s) presence	0.48	0.52	**0.50**	**0.68**	0.47	**0.56**	0.11	**0.63**	0.19
black frame	0.09	0.20	0.13	0.25	**0.63**	0.36	0.12	0.41	0.18
significant pause	0.32	0.34	0.33	0.21	**0.63**	0.32	0.02	0.38	0.04
transition phrase	**0.61**	0.32	0.42	0.44	0.55	0.49	0.20	0.47	**0.28**
bow histogram	0.30	**0.63**	0.40	0.26	0.54	0.35	0.11	0.47	0.19
reporter marker	0.20	**0.63**	0.30	0.34	0.28	0.31	**0.43**	0.25	0.31

4.1 Importance of Features

Before presenting the story boundary detection results using multimodal fusion, we compare the performance of individual features. Table 1 shows Precision and Recall values for the maximum F1 score of each feature, where F1 score is defined as $F1 = \frac{2 \cdot P \cdot R}{P+R}$. We observe a high F1 score for the anchor feature across all programs. From our observations, NBC and ABC News have more similar production rules. In these two programs, significant pause is an important feature where the reporter change marker is not a strong feature. These two programs tend to use >>> markers for story changes so the majority of the reporter change markers are not indicators of story change.

For PBS News Hour, reporter change and presence of transition phrase have high F1 scores where presence of black frame doesn't seem to be a strong feature. This is due to the fact that PBS News Hour does not have commercials. However, as we mentioned before, black frames may occasionally be inserted before story boundaries. Also, most of these stories start after transition music, which causes the pause feature to perform poorly.

From these results, we observe that the presence of each features is valuable. However, feature performance varies across news programs. Therefore, each channel requires a separate classifier and a different fusion method to achieve the best performance.

4.2 Multimodal Fusion

We evaluated story boundary detection over our database of three news programs. NBC Nightly News and ABC World News have >>> markers in their closed captioning text which indicate story changes. In these two programs, we do not use these error-prone story change markers as a feature for segmentation but we analyze their performance and consider them as a baseline for our segmentation performance analysis.

Figure 4 shows the Precision-Recall curves for NBC Nightly News and ABC World News. In NBC News, we observe that text features (T-F) alone are as good as the story change markers performance. Furthermore, both fusion techniques boost performance with late fusion (VA+T-LF) improving the baseline performance by more than 40%.

For ABC World News, video-aduio or text features alone perform better than the baseline. The best boundary detection performance for this program was an F1 score of 0.57 when using early fusion of video-audio and text features (VA+T-EF).

In Fig. 5 we show the Precision-Recall curve of PBS News Hour. This news program seems to benefit more from the text features, improving the video-audio feature based classification by more than 30% percent. The best F1 score is 0.41 for early-fusion of features (VA-EF).

From these results, we observe the importance of text features in improving segmentation accuracy across all programs. Furthermore, we believe there are many more useful text features we can extract from CC. Also, these results highlight the importance finding the proper fusion strategy for each channel for an optimal segmentation result.

5 Conclusion and Future Work

In this paper, we investigated different multimodal fusion methods for learning algorithms based on discriminative models. Along with video and audio cues, we utilized text cues from closed captioning which is readily available in all news streams. Furthermore, we proposed an empirical distribution-based feature representation for such features.

In addition, we investigated the importance of various modalities across three major U.S. news programs by analyzing the segmentation performance based on each modality. For each news program, we studied the importance of each

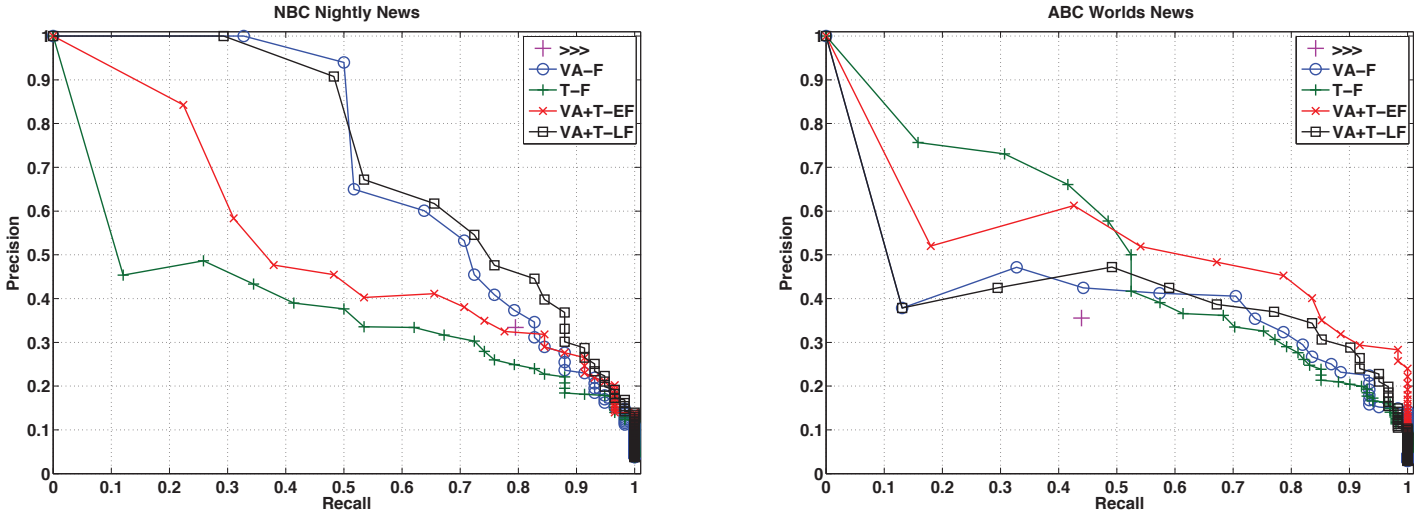

Fig. 4: Precision-Recall curve for NBC Nightly News (left), ABC World News (right)-comparing the performance of story change markers (>>>) text-features (T-F), video-audio features (VA-F), early fusion of all features (VA+T-EF) and late fusion of all features (VA+T-LF)

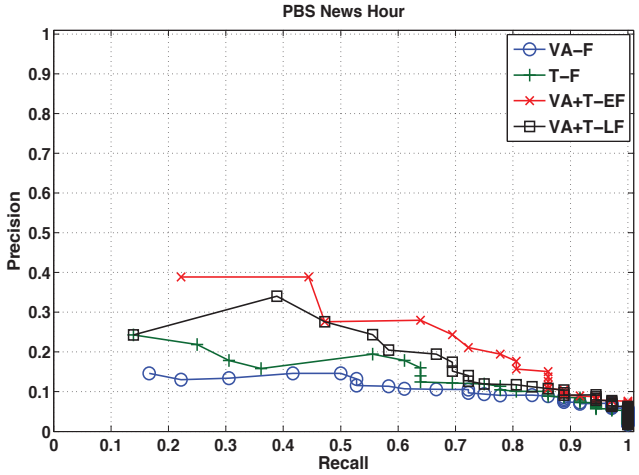

Fig. 5: Precision-Recall curve for PBS News Hour-comparing the performance of text-features (T-F), video-audio features (VA-F), early fusion of all features (VA+T-EF) and late fusion of all features (VA+T-LF)

feature and showed the need for per channel fusion strategy.

In the future, it is important to explore the temporal dynamics of news stories. The expected length of each story is a parameter than can be learned for each news program. Finally, major news programs produce hours of news per week following same production rules. It will be interesting to incorporate new sample data collected weekly without completely redesigning the classifier.

6 References

[1] W. Hsu, *An information-theoretic framework towards large-scale video structuring, threading, and retrieval*, Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University, 2007.

[2] Y. Zhai, A. Yilmaz, and M. Shah, "Story segmentation in news videos using visual and textual cues," in *ACM International Conference on Multimedia*, 2005, pp. 92–102.

[3] W. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *International Conference on Multimedia and Expo*, 2003, pp. 413–416.

[4] W. H. m. Hsu and S. F. Chang, "Generative, discriminative and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *IEEE International Conference on Multimedia and Expo*, 2004.

[5] 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 *SPIE Electronic Imaging*, 2004, pp. 244–258.

[6] W. Jianping, P. Tianqiang, and L. Bicheng, "News video story segmentation based on naive bayes model," in *International Conference on Natural Computation*, 2009, pp. 77–81.

[7] X. Gao and X. Tang, "Unsupervised and model-free news video segmentation," in *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2001, pp. 58–64.

[8] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[9] R. Y. D. Xu J. G. Allen and J. S. Jin, "Object tracking using camshift algorithm and multiple quantized feature spaces," in *Proceedings of the Pan-Sydney area workshop on Visual information processing*, 2004, pp. 3–7.

[10] A. G. Haupmann and M. J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Proceedings of the Advances in Digital Libraries Conference*, 1998, pp. 168–179.

[11] Pnar Duygulu, Ming yu Chen, and Er Hauptmann, "Comparison and combination of two novel commercial detection methods," in *Proceedings of the International Conference on Multimedia and Expo*, 2004, pp. 1267–1270.

[12] W. Walker, P. Lamere, P. P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: a flexible open source framework for speech recognition," Tech. Rep., 2004.

[13] C.W. Huang, W. Hsu, and S.-F. Chang, "Automatic closed caption alignment based on speech recognition transcripts," Tech. Rep., Columbia University, 2003.

[14] A. G. Hauptmann and M. J. Witbrock, "Story segmentation and detection of commercials in broadcast news vvideo," in *Advances in Digital Libraries Conference*, 1998, pp. 168–179.

[15] M. Worring C.G. M. Snoek and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM Multimedia*, 2005, pp. 399–402.

[16] G. Peter and N. Sebastian, "On feature combination for multiclass object classification," in *International Conference on Computer Vision*, 2009, pp. 221–228.