NEWS VIDEO STORY SEGMENTATION USING FUSION OF MULTI-LEVEL MULTI-MODAL FEATURES IN TRECVID 2003

W. Hsu[†], L. Kennedy[†], C.-W. Huang[†], S.-F. Chang[†], C.-Y. Lin[‡] and G. Iyengar[‡]

[†]Dept. of Electrical Engineering, Columbia University, NY
[‡]IBM T. J. Watson Research Center, NY

ABSTRACT

In this paper, we present our new results in news video story segmentation and classification in the context of TRECVID video retrieval benchmarking event 2003. We applied and extended the Maximum Entropy statistical model to effectively fuse diverse features from multiple levels and modalities, including visual, audio, and text. We have included various features such as motion, face, music/speech types, prosody, and high-level text segmentation information. The statistical fusion model is used to automatically discover relevant features contributing to the detection of story boundaries. One novel aspect of our method is the use of a feature wrapper to address different types of features – asynchronous, discrete, continuous and delta ones. We also developed several novel features related to prosody. Using the large news video set from the TRECVID 2003 benchmark, we demonstrate satisfactory performance (F1 measure up to 0.77) and more importantly observe an interesting opportunity for further improvement.

1. INTRODUCTION

News story segmentation is an important underlying technology for information exploitation in news video, which is a major information source in the new era. There have been several projects addressing news story segmentation, which could be categorized as heuristic rules or statistical approaches (see reviews in [1]). The former is mainly based on the assumption that each story starts with an anchor segment. Thus, the main theme of the work is to find the anchor segments with studio setup or anchor face/speech detection. These ad hoc algorithms lack the generality in handling diverse video sources with different features and production rules. The latter takes a statistical approach such as Hidden Markov Model. In our prior work [1], we adopt the Maximum Entropy (ME) approach by fusing dozens of features on hours of Mandarin news. In this work, we extend that approach by including novel perceptual features, solving multi-modal fusion issues with a novel feature wrapper and evaluating on 218 half-hour ABC/CNN news programs.

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. Other coherent segments are labelled as non-news. These non-news stories cover a mixture of footage: commercials, lead-ins, and reporter chit-chat. A story can be composed of multiple shots; e.g., an anchorperson introduces a reporter and the story is finished back in the studio-setting. On the other hand, a single shot can contain multiple story boundaries; e.g., an anchorperson switching to the next news topic. We excerpt some of the common story types in Figure 1. To assess the baseline performance, we also conduct an experiment by evaluating story boundaries with visual anchor segments only and yield a baseline result, shown in Table 1, where boundary detection F1[1] measures in ABC is 0.67 and is 0.51 in CNN with only 0.38 recall and 0.80 precision rates. The definition of evaluation metrics is explained in Section 4.1. In this paper, we will present significant performance gain over the baseline by using statistical multi-modal fusion

In TRECVID 2003, we have to detect all the boundaries at the transition from news to another news, news to non-news, and non-news to news segments. Furthermore, we label segments between boundaries as "news" or "non-news".

The issues regarding multi-modal fusion are discussed in Section 1.2. The probabilistic framework and the feature wrapper are addressed in Section 2. Relevant features are presented in Section 3. The experiment, evaluation metrics, and discussions are listed in Section 4 and followed by the conclusion and future work in Section 5.

1.1. Data set

In this work, we use 218 half-hour ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January 1998 through June 1998. The video is in MPEG-1 format and with associated files including automatic speech recognition (ASR) transcripts and annotated story boundaries, called *reference boundaries*. The data are prepared for TRECVID 2003[2] with the goal to promote progress in content-based video retrieval via open metric-based evaluation.

1.2. Issues with multi-modal fusion

There are generally two perspectives on story segmentation – one is boundary based and the other is segment based. The former models the characteristics of features at the boundary points; the latter models the temporal dynamics within each story. We adopt the first approach in this paper. In such an approach, one basic issue is the determination of candidate points each of which is tested and classified to story boundary or not.

Candidate points: A good candidate set should have a very high recall rate on the reference boundaries and are the places where salient and effective features occur. They are usually shot boundaries in most news segmentation projects (see reviews in [1]). However, we found that taking the shot boundaries only is

Thanks to Martin Franz of IBM Research for providing an ASR only story segmentation system.

[1] $F1 = \frac{2 \cdot P \cdot R}{P+R}$, where P and R are precision and recall rates

[2] TRECVID 2003: http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

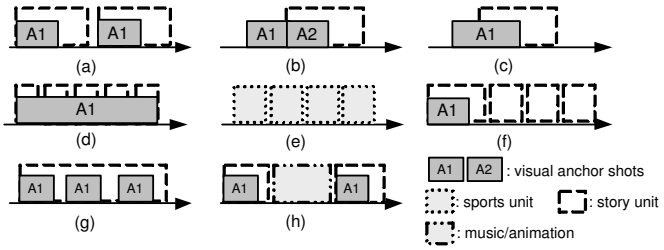

Fig. 1. Common story types seen in the ABC/CNN programs: (a) two normal stories start with anchor person; (b) a story starts after switching to a different visual anchor; (c, d) stories start within anchor shots; (e) a sports section constitutes series of briefings; (f) series of stories do not start with anchor shots; (g) multiple anchor shots appear in the same story unit; (h) two stories are separated by long music or animation representing the station id.

not complete. We evaluate the candidate completeness by detecting reference boundaries with 5-second fuzzy window (defined in Section 4.1). Surprisingly the recall rate for ABC/CNN on the shot boundaries is only 0.91. The reason is that some reference boundaries are not necessarily at the shot boundaries. In this work, we take the union of shot boundaries and audio pauses as candidate points but remove duplications within a 2.5-second fuzzy window. The union candidates yield 100% recall rate.

Data labelling: We adopt a supervised learning process with manually annotated reference boundaries. Since the features are usually asynchronous across modalities and the annotated data are not necessarily aligned well with the ground truth, each candidate point is labelled as "1" (boundary) if there is a reference boundary within the 2.5-second fuzzy window. However, some reference boundaries could not locate corresponding candidates within the fuzzy window. The phenomenon also happens in our ASR text segmentation and we just insert these reference boundaries as additional candidate points in the training set.

2. PROBABILISTIC FRAMEWORK

News videos from different channels usually have different production rules or dynamics. We choose to construct a model that adapts to each different channel. When dealing with videos from unknown sources, identification of the source channel can be done through logo detection or calculating model likelihood (fitness) with individual statistical station models.

We propose to model the diverse production patterns and content dynamics by using statistical frameworks. The assumption is that there exist consistent statistical characteristics within news video of each channel, and with adequate learning, a general model with a generic pool of computable features can be systematically optimized to construct effective segmentation tools for each news channel. We summarize the model and processes in this section, leaving details in [2].

2.1. Maximum Entropy model

The ME model [1, 3] constructs an exponential log-linear function that fuses multiple features to approximate the posterior probability of an event (i.e., story boundary) given the audio, visual or text data surrounding the point under examination, as shown in Equation 1. The construction process includes two main steps - parameter estimation and feature induction.

The estimated model, a posterior probability, is represented as $q_\lambda(b|x)$, where $b \in \{0, 1\}$ is a random variable corresponding to the presence or absence of a story boundary in the context x and λ is the estimated parameter set. Here x is the video and audio data surrounding a candidate point of story boundaries. From x we compute a set of binary features, $f_i(x, b) = 1_{\{g_i(x)=b\}} \in \{0, 1\}$. $1_{\{\cdot\}}$ is a indication function; g_i is a predictor of story boundary using the i'th binary feature, generated from the feature wrapper (Section 2.2). f_i equals 1 if the prediction of predictor g_i equals b, and is 0 otherwise.

Given a labelled training set, we construct a linear exponential function as the following:

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} \exp\{\sum_i \lambda_i f_i(x, b)\}, \tag{1}$$

where $\sum_i \lambda_i f_i(x, b)$ is a linear combination of binary features with real-valued parameters λ_i. $Z_\lambda(x)$ is a normalization factor to ensure Equation 1 is a valid conditional probability distribution. Basically, λ_i controls the weighting of i'th feature in estimating the posterior probability.

Parameter estimation: The parameters $\{\lambda_i\}$ are estimated by minimizing the Kullback-Leibler divergence measure computed from the training set that has empirical distribution $\tilde{p}$. The optimally estimated parameters are

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, D(\tilde{p} \parallel q_\lambda), \tag{2}$$

where $D(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence.

Feature induction: Given a set of prospective binary features $\mathbf{C}$ and an initial maximum entropy model q, the model can be improved into $q_{\alpha,g}$ by adding a new feature $g \in C$ with a suitable weight α, represented as

$$q_{\alpha,g}(b|x) = \frac{\exp\{\alpha g(x, b)\} q(b|x)}{Z_\alpha(x)}, \tag{3}$$

where $Z_\alpha(x)$ is the normalization factor. A greedy induction process is used to select the feature that has the largest improvement in terms of gains or divergence reduction. The iteration process repeats till stopping criterion is reached (e.g., upper bound of the number of features or lower bound of the gain).

Generalization error: The divergence optimization is actually convex. In each feature induction, thus, the selection process could always pick up a feature that induces gains to the current model. However, the continuing iterations would overfit the training set. We modify the stopping criterion of feature induction with the time when the iteratively improved model on the training set has a degraded F1 measure on a sperate validation set, comparing with the previous constructed model. This criterion is used to avoid overfitting.

2.2. Feature wrapper

In Figure 2, we show the relation between the feature wrapper and the feature library which stores all raw multi-modal features. As the raw feature f_i^r is taken into the feature wrapper, it will be rendered into sets of binary features at each candidate point $\{t_k\}$ with the function $F_w(f_i^r, t_k, dt, v, B)$, which is used to take features from observation windows of various lengths B, compute delta values of some features over time interval dt, and finally binarize the feature values against multiple possible thresholds, v.

Delta feature: The delta feature is quite important in human perception according to our experiment; for example, the motion

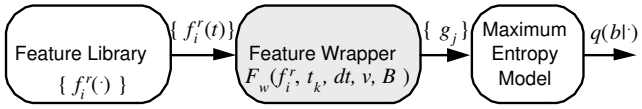

Fig. 2. The raw multi-modal features f_i^r are collected in the feature library and indexed by raw feature id i and time t. The raw features are further wrapped in the feature wrapper to generate sets of binary features $\{g_j\}$, in terms of different observation windows, delta operations, and binarization threshold levels. The binary features are further fed into the ME model.

intensity drops directly from high to low. Here we get the delta raw features by comparing the raw features with the time difference dt as $\Delta f_i^r(t) = f_i^r(t) - f_i^r(t - dt)$. Some computed delta features, in real values, will be further binarized in the binarization step.

Binarization: The story transitions are usually correlated with the changes in some dominant features near the boundary point. However, there are no prior knowledge about the quantitative threshold values for us to accurately detect "significant changes". For example, what is the right threshold for the pitch jump intensity? How far would a commercial starting point affect the occurrence of a story boundary? Our strategy should be to find the effective binarization threshold level in terms of the fitness gain (i.e., divergence reduction defined in Equation 2) of the constructed model rather than the data distribution within the feature itself. Each raw or delta feature is binarized into binary features with different threshold levels v.

Observation windows: Different observation windows also impact human perception on temporal events. Here we take three observation windows B around each candidate t_k. The first is the interval before the candidate point with window size T_w; the next is the same time-span after the candidate; the other is the window surrounding the candidate, $[t_k - T_w/2, t_k + T_w/2]$. With different observation windows, we try to catch effective features occurring before, after, or surrounding the candidate points. This mechanism also tolerates time offset between different modalities. For example, the text segmentation boundaries or prosody features might imply likely occurrence of true story boundaries near a local neighborhood but not a precise location.

The dimension of binary features $\{g_j^i\}$ generated from raw feature f_i^r or delta feature Δf_i^r is the product of the number of threshold levels and number of observation windows (3, in our experiment). All the binary features generated at a candidate point are sequently collected into $\{g_j\}$ and are further fed into the ME model; e.g., for pitch jump raw feature with 4 threshold levels, it would generate $3 \cdot 4 = 12$ binary features since we have to check if the feature is "on" in the 3 observation windows and each is binarized with 4 different levels.

2.3. Segment classification

In TRECVID 2003, another task is to classify the detected video segment to "news" vs. "non-news". Although sophisticated models can be built to capture the dynamics and features in different classes, we adopt a simple approach so far. We apply a separate commercial detector (described below) to each shot and simply compute the overlap between the computed boundary segments and the detected commercial segments. The computed segment is labelled as news if it overlaps the non-commercial portions more than a threshold; otherwise is labelled as non-news. The threshold is determined from the training set with the best argument that maximizes story classification F1 measure.

The intuition is that boundary detection might be erroneous but we could still classify the segment by checking the surrounding context, commercial or non-commercial. This simple approach will make mistakes for shorts segments such as chit-chat, station animations, which are not commercials but should be classified as non-news. However, such errors may not be significant as the percentage of such anomaly segments is usually small.

3. RAW MULTI-MODAL FEATURES

The raw features, reposited in feature library as shown in Figure 2, are from different feature detectors. Due to space limitations, some of them are presented below and more details are in [2].

Anchor face: The anchor face detector is as [1] but enhanced with another detector using GMM skin-tone model and geometric active contour to locate the possible face regions [2].

Commercial: Matching based on image templates such as station logos and caption titles is used to discriminate commercial and non-commercial frames. After the template detection, morphological operators are applied to smooth the result with temporal consistence.

Pitch jump: Pitch contour has been shown to be a salient feature for the detection of syntactically meaningful phrase and topic boundaries [4, 5] and independent of language and gender [4]. Different from those methods, we invent a new algorithm by converting pitch estimates into octaves, which are further normalized with mean pitch in same-speaker segments from the ASR output. The pitch jump points are found by searching for points in the speech where the normalized magnitude of the inter-chunk pitch change is above a certain normalized threshold, where a chunk is a group of adjacent valid pitch estimates.

Significant pause: Significant pause is another novel feature that we develop for the news video segmentation task. It is inspired by the "pitch reset" behavior [5] in the pitch jump feature and the "significant phrase" feature developed in [6]. Significant pause is essentially an AND operation of pitch jump points and pauses. We look for coincidences of pitch jump and pause time points in an attempt to capture the behavior where news anchors may simultaneously take a pause and reset their pitch contour between news stories. To gauge the potential significance of this feature before fusing into the framework, we test it on the reference boundaries and found its F1 measure to be 0.42, which is quite impressive compared with other features. As a comparison, another salient feature, anchor face, contributes to the story boundaries with a F1 measure of 0.51.

Speech segments and rapidity: We extract the speech segment, a continuous segment of the same speaker, from the ASR outputs. Two adjacent speech segments might belong to the same speaker since being separated by a long non-speech segments, pauses or music. The segment boundary might imply a story boundary. We further measure the speech rapidity by counting words per second in each segment.

ASR-based story segmentation: The ASR-based story segmentation scheme is a combination of decision tree and maximum entropy models. It takes a variety of lexical, semantic and structural features as inputs and generates boundary scores at non-speech candidates, where no ASR words are transcribed [7].

Combinatorial features: Some combinatorial features from other detectors are fused to highlight some rare events such as pitch jump points near the starts of the speech segments, significant pauses within the fast speech segments, and significant pauses

Table 1. Boundary detection performance in ABC/CNN news.

	ABC			CNN		
Modalities	P	R	$F1$	P	R	$F1$
Anchor Face	0.67	0.67	0.67	0.80	0.38	0.51
T	0.65	0.55	0.59	0.50	0.70	0.59
A+V	0.75	0.67	0.71	0.80	0.54	0.65
A+V+T	0.88	0.68	0.77	0.83	0.58	0.68

near shot boundaries.

4. EXPERIMENTS

In this experiment, we use 111 half-hour video programs for development, 66 of which are used for detector training and threshold determination. The remaining 45 video programs are further separated for fusion training and model validation. In TRECVID 2003, we have to submit the performance of a sperate test set, composed of 107 ABC/CNN videos. The performance in the submission is similar to what we obtain in the validation set except that the submitted CNN recall rate is slightly lower. Here we present the performance of the evaluations from the validation set only.

4.1. Boundary detection performance

The segmentation measure metrics are precision P_{seg} and recall R_{seg} and are defined as the following. According to the TRECVID 2003 metrics, each reference boundary is expanded with a fuzzy window of 5 seconds in each direction, resulting in an evaluation interval of 10 seconds. A reference boundary is *detected* when one or more computed story boundaries lie within its evaluation period. If a computed boundary does not fall in the evaluation interval of a reference boundary, it is considered a *false alarm*. The precision P_{seg} and recall R_{seg} are defined in Equations 4 and 5, where $|\cdot|$ means the number of boundaries.

$$P_{seg} = \frac{|\text{computed boundaries}| - |\text{false alarms}|}{|\text{computed boundaries}|} \quad (4)$$

$$R_{seg} = \frac{|\text{detected reference boundaries}|}{|\text{reference boundaries}|} \quad (5)$$

The performance in the development set is shown in Table 1, where "A" means audio cues, "V" is visual cues and "T" is text. At A+V, the recall rate of ABC is better than CNN; however, the precision is somehow lower. It is probably due to that ABC stories are dominated by anchor segments or types (a) and (g) in Figure 1; while in CNN, there are some short briefings and tiny dynamic sports sections which are very challenging and thus cause a lower recall rate. About CNN news, the A+V boosts the recall rate of anchor face from 0.38 to 0.54 and does not degrade the precision. The main contributions come from significant pauses and speech segments since they compensate CNN's lack of strong visual cues.

As for fusing modality features such as fusing text segmentation into A+V, the precision and recall are both improved even though the text feature is with real-valued scores and computed at non-speech points only, which may not coincide with those used for the audio-visual features. It is apparent that the fusion framework successfully integrates these heterogeneous features which compensate each other.

We try to ensure that we have adequate training sample size. For example, to train a CNN boundary detection model with A+V modalities, we use 34 CNN videos ($\sim$17 hours) with 1142 reference boundaries and 11705 candidate points. Each candidate is with 186 binary features, among which the feature induction process selects 30 of them.

4.2. Segment classification performance

Each detected segment is further classified into news vs non-news using the algorithm described above. We observe high accuracy of segment classification (about 0.91 in F1 measure) in both CNN and ABC. Similar accuracies are found in using different modality fusions, either A+V or A+V+T. Such invariance over modalities and channels is likely due to the consistently high accuracy of our commercial detector.

5. CONCLUSION AND FUTURE WORK

Story segmentation in news video remain a challenging issue even after years of research. We believe multi-modality fusion through effective statistical modelling and feature selection are keys to solutions. In this paper, we have proposed a systematic framework for fusing multi-modal features at different levels. We demonstrated significant performance improvement over single modality solutions and illustrated the ease in adding new features through the use of a novel feature wrapper and ME model.

There are other perceptual features that might improve this work; for example, an inter-chunk energy variations might be highly correlated with the pitch reset feature discussed earlier; another one is the more precise speech rapidity measured at the phoneme level since towards the end of news stories news anchors may have the tendency to decrease their rate of speech or stretching out the last few words. In addition, the cue terms extracted from embedded text on the image might provide important hints for story boundary detection as well.

According to our observation, a ME model extended with temporal states would be a promising solution since the statistical behaviors of features in relation to the story transition dynamics may change over time in the course of a news program.

6. REFERENCES

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

[2] W. Hsu and S.-F. Chang, "Discovery and fusion of salient multi-modal features towards news story segmentation," Tech. Rep., Columbia University, October 20 2003.

[3] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, no. special issue on Natural Language Learning, pp. 177–210, 1999.

[4] J. Vaissiere, "Language-independent prosodic features," in *Prosody: Models and Measurements*, Anne Cutler and D. Robert Ladd, Eds., pp. 53–66. Springer, Berlin, 1983.

[5] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[6] H. Sundaram, *Segmentation, Structure Detection and Summarization of Multimedia Sequences*, Ph.D. thesis, Columbia University, 2002.

[7] 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.