A STATISTICAL FRAMEWORK FOR FUSING MID-LEVEL PERCEPTUAL FEATURES IN NEWS STORY SEGMENTATION

Winston H.-M. Hsu, Shih-Fu Chang

Department of Electrical Engineering, Columbia University
{winston, sfchang}@ee.columbia.edu

ABSTRACT

News story segmentation is essential for video indexing, summarization and intelligence exploitation. In this paper, we present a general statistical framework, called *exponential model* or *maximum entropy model*, that can systematically select the most significant mid-level features of various types (visual, audio, and semantic) and learn the optimal ways in fusing their combinations in story segmentation. The model utilizes a family of weighted, exponential functions to account for the contributions from different features. The Kullbak-Leibler divergence measure is used in an optimization procedure to iteratively estimate the model parameters, and automatically select the optimal features. The framework is scalable in incorporating new features and adapting to new domains and also discovers how these feature sets contribute to the segmentation work. When tested on foreign news programs, the proposed techniques achieve significant performance improvement over prior work using ad hoc algorithms and slightly better gain over the state of the art using HMM-based models.

1. INTRODUCTION

News video represents a major source of information in the new era of the information society. Intelligent systems are needed to summarize, categorize, personalize and even discover relations between relevant news stories. To support these functionalities, there are many works addressing an important underlying technology, news story segmentation. A well-known project is the *Informedia* project, in which rules on combining image, audio, transcripts and closed captions are used to locate story boundaries [1]. However, for international news programs, closed captions and accurate speech recognizers are usually unavailable. Besides, the production rules vary from different channels or countries. In [2], Qi, *et al*, identify the story boundaries by a clustering based algorithm that detects anchorpersons by performing an AND operation on anchorperson shots and anchorperson speech segments. The image clustering method is applied on the entire key frames and may not achieve required accuracy. Liu, *et al* propose a similar method and construct online visual and acoustic cues to identify anchorpersons [3]. Their face region detection process is applied to key frame of each shot only and is sensitive to shot detection error. In our prior work [4], we have developed a real-time story segmentation system using compressed-domain face detection, anchorperson shot detection, and clustering using face-region features and speech features. The system achieves real-time efficiency and reasonable accuracy, but suffers from the use of an ad hoc algorithm and the lack of systematic framework to handle news from diverse sources. A promising statistic framework based on HMM is presented in [5]. The authors employ a decision tree to classify shots into one of the 13 pre-defined categories and then category labels are fed to a Hidden Markov Model (HMM) to locate the story boundary.

Most news story systems use ad hoc algorithms and lack the generality in handling diverse sources with different features and production rules. Besides, all these approaches work on English news programs only.

In this paper, we present a new statistical framework for news video story segmentation that is scalable to different sources and production rules. We base on the approach of effective detection of various mid-level features from audio-visual data, and a robust statistic framework in systematically learning the rules of fusing mid-level features. We extend *the exponential model* (also called *maximum entropy model*) used in [7][8] for topic segmentation in text documents, where a family of weighted, exponential functions of binary features is used to account for the text boundary likelihood at each sentence boundary. Such features may include the occurrences of certain key terms, such as *Does the word MR appear in the next sentence?*, *Does the word SAID appear in the previous five sentences but not in the next five sentences?*. The authors also use an induction procedure to automatically find the most salient features. We extend the ideas to the audio-visual data in news programs, in which mid-level features like occurrence of anchorperson, speaker/audio change, superimposed text caption changes, cue phrases, etc., contribute to classifying candidate points to story boundaries.

The adopted statistical model is scalable to various news sources in that different mid-level features can be easily added to the family of exponential functions. For international news programs that do not have closed captions and reliable speech recognition, the framework offers an effective solution that other audio-visual cues can be readily integrated. We have tested the approach on foreign news (Taiwanese news) and achieved a promising performance – 90% F1 measure, 87% precision, and 94% recall. This is significantly better than the ad hoc approach we developed in [4] and slightly better than the top performance reported in the literature [5] using a different statistical framework based on HMM decoding of mid-level features[*]. Another unique benefit of the framework is its efficiency in selecting the most significant features in decoding story boundaries through a quantitative modeling and computation of a probabilistic divergence measure. Our feature

[*] Note the comparison is not based on standard benchmark data and procedures, numerical performance difference shouldn't be emphasized.

induction results confirm the popular assumption that anchorperson shot is a good indicator of story boundary, but it must be enhanced with other features such as cue phrases to achieve a significantly better accuracy. The observation is shown in section 4.3.

We present the exponential model framework and feature induction procedure in Section 2. In Section 3, we describe a dynamic programming technique to incorporate the information about story length distribution. Experiments and discussions of significant features are reported in Section 4.

2. EXPONENTIAL MODEL

The exponential model 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 data surrounding the point under examination, as equation (1). The construction process includes two main steps – parameter estimation and feature induction.

The exponential posterior probability is denoted as $q(b \mid x)$, where $b \in \{0,1\}$ is a random variable corresponding to the presence or absence of a story boundary in the context x. Here x is the video and audio data surrounding a candidate point of story boundaries. The candidate points can be chosen in various ways, such as all the shot change points or the boundaries of the significant pauses in the speech. We currently use the shot boundaries in our work. From x we compute a set of binary variables, $f_i(x,b) = 1_{\{g_i(x)=b\}} \in \{0,1\}$, $i = 1...K$. $1_{\{\cdot\}}$ is an indication function; g_i is a predictor of story boundary using the i'th feature, such as occurrence of anchorperson, occurrence of cue phrases, superimposed text change, etc. f_i equals 1 if the prediction of predictor g_i equals b, and is 0 otherwise; K is the total number of predictors.

Given a training data set $\tilde{D} = \{(x_j, b_j)\}$, $j = 1...K$, we construct a linear exponential function as follows

$$q(b \mid x) = \frac{1}{Z_l(x)} e^{\sum_i l_i \cdot f_i(x,b)}, \quad (1)$$

where $\sum_i l_i f_i(x,b)$ is a linear combination of binary features with real-valued parameters l_i. $Z_l(x) = e^{\sum_i l_i \cdot f_i(x,0)} + e^{\sum_i l_i \cdot f_i(x,1)}$ is a normalization factor to ensure equation (1) is a valid conditional probability distribution. Basically, l_i controls the contribution of i'th feature in estimating the posterior probability of the current candidate point being a story boundary.

2.1. Estimation of Parameter $\{l_i\}$

The parameters $\{l_i\}$ are estimated by minimizing Kullbak-Leibler divergence measure computed from $\tilde{D}$ that has empirical distribution $\tilde{p}(b,x)$. Since each video chunk is different we just let $\tilde{p}(b,x) = 1/|\tilde{D}|$, where $|\tilde{D}|$ is the number of samples in $\tilde{D}$. The optimal estimation is

$$q_l^* = \arg\min_l D(\tilde{p} \,\|\, q_l), \quad (2)$$

and $D(\cdot \,\|\, \cdot)$ is the divergence defined as:

$$D(\tilde{p} \,\|\, q_l) = \sum_x \tilde{p}(x) \sum_{b \in \{1,0\}} \tilde{p}(b \mid x) \log \frac{\tilde{p}(b \mid x)}{q_l(b \mid x)}. \quad (3)$$

In [8], it is shown that minimizing the divergence is equivalent to maximizing the log-likelihood defined as

$$L_{\tilde{p}}(q_l) \equiv \sum_x \sum_b \tilde{p}(x,b) \log q_l(b \mid x). \quad (4)$$

Thus the exponential model is also called *maximum entropy* or *minimum divergence*. The log-likelihood is used to measure the quality of the model and $L_{\tilde{p}}(q_l) \le 0$ holds all the time. In the ideal case, $L_{\tilde{p}}(q_l) = 0$ corresponds to a model q_l, which is "perfect" with respect to $\tilde{p}$; that is, $q_l(b \mid x) = 1$ if and only if $\tilde{p}(x,b) > 0$.

We use an iterative process to update $\{l_i\}$ till divergence is minimized. In each iteration, $l_i' = l_i + \Delta l_i$, where

$$\Delta l_i = \frac{1}{M} \log \left(\frac{\sum_{x,b} \tilde{p}(x,b) f_i(x,b)}{\sum_{x,b} \tilde{p}(x) q_l(b \mid x) f_i(x,b)} \right), \quad (5)$$

and M is a constant to control the convergence speed. This formula updates the model in a way so that the expectation values of features f_i with respect to the model are the same as their expectation values with respect to the distribution from the training data. In other words, the expected fraction of events (x,b) for which f_i is "on" should be the same no matter it's measured based on the empirical distribution $\tilde{p}(x,b)$ or the estimated model $\tilde{p}(x) q_l(b \mid x)$. When the exponential model underestimates the expectation value of feature f_i, its weight l_i is increased. Conversely, l_i is decreased when the overestimation occurs.

2.2. Feature induction

Given a set of candidate features C and an initial exponential model q, the model can be refined by adding a new feature $g \in C$ with weight a.

$$q_{a,g}(b \mid x) = \frac{e^{a\,g(x,b)} q(b \mid x)}{Z_a(x)}, \quad (6)$$

where $Z_a(x) = \sum_{b \in \{0,1\}} e^{a\,g(x,b)} q(b \mid x)$ is the normalization factor. A greedy induction process is to select the feature that has the largest improvement in terms of gain $G_q(g)$, or divergence reduction and

$$G_q(g) = \sup_a \left(D(\tilde{p} \,\|\, q) - D(\tilde{p} \,\|\, q_{a,g}) \right) = \arg\max_a G_{qg}(a), \quad (7)$$

$$G_{qg}(a) \equiv L_{\tilde{p}}(q_{a,g}(b \mid x)) - L_{\tilde{p}}(q(b \mid x)) = -\sum_X \tilde{p}(x) \log Z_a(x) + a \sum_x \sum_b \tilde{p}(x,b) g(x,b). \quad (8)$$

After calculating the gains associated with all the candidates for new features, the one with the highest gain is added to the model. The iteration process repeats till the desired size of model (i.e., the number of features) is achieved or a satisfactory value of divergence is reached. During feature induction, the gain $G_{qg}(a)$ is calculated according to equation (8). Each feature g is computed with the optimal weight a under the assumption that the other parameters remain fixed. Since the optimal a occurs when $G'_{qg}(a) = 0$, we could apply the Newton's method on $G'_{qg}(a)$ to locate the optimal a. The iterative step is defined as $a_{n+1} = a_n - \frac{G'_{qg}(a)}{G''_{qg}(a)}$.

3. DYNAMIC PROGAMMING

Typically the length of news story is not random; instead, it follows a probability distribution that can be estimated from some training data. Such distribution is affected by the production rules used in news sources. Given the estimated distribution of story length, we apply a dynamic programming approach to smooth the likelihood output from the above exponential model $q_l(b \mid x)$. We use a random variable X_s to represent the story length and its correspondent cumulative distribution function is $P_s(x) = \Pr\{X_s \le x\}$. Estimation of such a distribution is based on the empirical data in the training set. We further model the transition probability $a_{ij}^{(k)}, i, j \in \{1,0\}$, at each hypothesis point k, with time index t_k. This represents the conditional probability that the current point at time t_k is (or is not) a story boundary given the previous shot boundary at time t_{k-1} is (or is not) a story boundary, and the most recent story boundary occurs at time t_{k^-}. Such transition probabilities can be computed from the accumulative distribution function mentioned above. Details are skipped here due to space constraints [4].

The recursion processes based on the dynamic programming method are defined as the following,

$$\tilde{d}_k(j) = \max_{i \in \{1,0\}} \left[\tilde{d}_{k-1}(i) + \tilde{a}_{ij}^k \right] + \tilde{q}(j \mid x_k) \quad (9)$$

$$y_k(j) = \arg\max_{i \in \{1,0\}} \left[\tilde{d}_{k-1}(i) + \tilde{a}_{ij}^k \right], \mathrm{j} \in \{1, 0\} \quad (10)$$

where $\tilde{a}_{ij}^k = \log(a_{ij}^k)$ and $\tilde{q}(j \mid x_k) = \log(q(j \mid x_k))$. $\tilde{d}_k$ is the accumulative likelihood (in log form) for class j at time index k and $y_k(j)$ is the optimal class selected at time t_{k-1} through backtracking for class j at time t_k.

4. EXPERIMENTS

4.1. Mid-Level Features

We focus on mid-level features of various types including visual, acoustic, and semantic ones. Given a candidate point, i.e., a shot boundary, we compute the mid-level features from the audio-video data surrounding the candidate point. The features are thresholded into binary values. We consider the following categories of features.

Acoustics: Four binary features: (1) if the next shot is dominated by speech, (2) if the next shot starts with music, (3) if the previous shot is dominated by speech, and (4) if the previous shot ends with music. These four features are detected by using a music/speech discrimination package [4] with success rate around 85%.

Speaker Identification: We use the anchorperson shot detector and the speech matching module from our prior work [4] to decide if the following shot is an anchor speech segment.

Face: Two binary features: (1) whether the previous shot has a static face and (2) whether the following shot has a static face. We use the static face detector we developed in [4].

Superimposed text captions: We used our prior results in [6] to compute three features: (1) if the next shot has a caption, (2) if the previous shot has a caption, and (3) if both neighboring shots have the same caption.

Motion: Two features: (1) if the next shot is with low motion and (2) if there is motion intensity change (low to high or high to low) from the previous shot to the next shot.

A/V Combination: Two combinatorial features by inspecting next shot with (1) AND(with a static face, dominated with speech) and (2) AND(with a static face, anchor speech segment).

Cue Phrase: We check whether any cue phrases from a manually selected set exist in the surrounding speech. Such cue phrases are closer to the semantic level than the above audio-visual features, and are useful indicators of story boundaries. For Mandarin speech, we adopt the syllable-level combinations suggested in [9][†] to define the cue phrases. We select ten frequent cue phrases at the end and beginning of story boundaries in the training set. The syllable-level combinations have been shown effective in broadcast speech information retrieval.

4.2. Performance

We deliberately select foreign news (Mandarin news in Taiwan) to test our model and foreign news programs usually do not have special markers in closed captions to indicate the story boundaries. We collect a data set of 3.5-hour news programs with totally 100 stories, mixing from 3 different channels and different times. The corpora are with 8 unique anchorpersons in 10 video clips. We use a subset (38 stories) for training and the rest for testing. We have implemented all the proposed components, including model parameter estimation, feature induction, and dynamic programming smoothing based on length distribution. For feature selection, we use the induction process to automatically select eight best features from different combinations of feature categories. The results are shown in table 1. It is not surprising to see that when we incorporate all categories of features, we obtain the best performance (precision 0.87, recall 0.94, F1 0.9).

Sets	CB	FR	FA	R	P	F1

[†] Note the complete process of speech recognition is not attempted here due to limited performance. Instead, syllable-level cue phrase detection is used. We use the Chinese speech recognition tools kindly provided by the authors of [9].

S+C+A+V	58	4	9	0.94	0.87	0.90
S+A+V	56	6	8	0.90	0.88	0.89
A+V+C	53	9	14	0.85	0.79	0.82
A+V	54	8	18	0.87	0.75	0.81
S	26	36	22	0.42	0.54	0.47

Table 1: Story segmentation accuracy when eight features are automatically selected from different combinations of feature categories. *S*={cue phrase}, *A*={acoustics ∪ speaker identification}, *V*={face ∪ superimposed caption ∪ motion} and *C*={A/V combination}. E.g. "*S+A+V*" includes feature categories *S*, *A* and *V*. CB: # of correct boundaries, FR: false rejection, FA: false alarm, R: recall, P: precision, F1: F1 measure combining P and R.

4.3. Significant features in story segmentation

As described in Section 2, the optimal features are selected based on their effects on divergence reduction. Log-likelihood, $L_{\tilde{p}}(q_l)$, in equation (4) is used to measure the quality of the estimated model q_l. Given the current training data, the ideal model should have a log-likelihood of 0 and a random-guess model has -0.6931. The log-likelihood value after each feature selection is shown and compared in Figure 1. Two equally good models are from the feature categories *S+C+A+V* and S+A+V with log-likelihood close to 0. A notable degradation in log-likelihood is observed when the {cue phrase} (S) category is excluded. However, when the {cue phrase} (S) category is used alone, the performance is degraded significantly. This indicates that the {cue phrase} features are helpful when combined with the audio-visual features. But when used alone, the audio-visual combinations are better than the {cue phrases} features.

We further inspect the first best feature selected in each feature set. In *S+C+A+V* and *A+V+C*, the first feature is an A/V combination feature, i.e., AND(static face, speaker speech segment). In the *A+V* set, the first feature selected is anchor speech segment and the second feature is static face. These findings confirm the popular assumption that story boundaries usually start with the anchorperson shot. Surprisingly, in *S+A+V* the induction process did not pick the above A/V feature as the first selection, instead, a specific type of cue phrase near the end of a story is picked. This difference may be due to that the static face/speaker speech feature may not always correspond to true story boundary, e.g., shots of interview or with the background color similar to skin-tone. In [4], we found the anchorperson model alone achieves only a limited performance (F1=0.65).

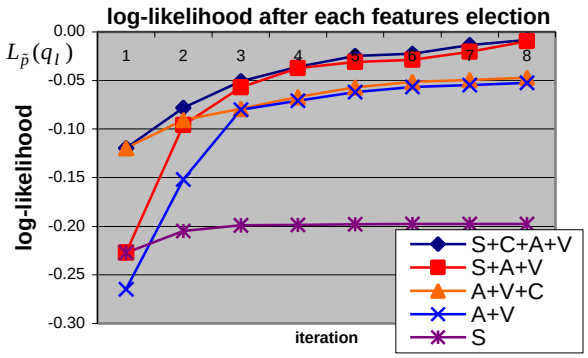

Figure 1: Log-likelihood $L_{\tilde{p}}(q_l)$ after each feature selection. The ideal value is 0 and the random-guess log-likelihood on current training data is –0.6931.

5. CONLUSIONS AND FUTURE WORKS

We have developed a probabilistic framework that systematically fuses multiple mid-level perceptual features, including visual, audio and semantic cues, to achieve a promising accuracy in news story segmentation (0.90 F1 measure, 0.87 precision, and 0.94 recall). We also demonstrated the capability in automatic selection of salient features by measuring their probabilistic divergence reduction. We are now expanding the acoustic feature sets, revising the cue phrase set, and changing the hypothesis points from shot boundaries to significant speech pauses to further improve the effectiveness of the system. In addition, we are applying the framework to news programs from different countries to analyze the commonality and difference of performance, salient features and production rules among diverse sources.

6. REFERENCES

[1] A. G. Hauptmann, M. J. Witbrock, "Story Segmentation and Detection of Commercials in Broadcast News Video", ADL-98 advances in Digital Libraries Conference, Santa Barbara, CA, April 22-24, 1998.

[2] W. Qi, Li. Gu, H. Jiang, X. R. Chen and H. J. Zhang, "Integrating Visual, Audio and Text Analysis for News Video," Proc. of ICIP, September 2000.

[3] Z. Liu and Q. Huang, "Adaptive Anchor Detection Using On-line Trained Audio/Visual Model," Proc. of SPIE, January 2000.

[4] W. Hsu, S.-F. Chang, "Frameworks for Fusing Mid-level Perceptual Features in News Story Segmentation," Columbia University ADVENT Technical Report, December 2002.

[5] L. Chaisorn, T. S. Chua and C. H. Lee, "The Segmentation of News Video into Story Units," Proc. of ICME, August 2002.

[6] D. Zhang, R. K. Rajendran, and S.-F. Chang, "General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video," Proc. of ICIP, September 2002.

[7] D. Beeferman and A. Berger and J. D. Lafferty, "Statistical Models for Text Segmentation," Machine Learning, p. 177-210, vol. 34, 1999.

[8] A. Berger, S. D. Pietra, and V. Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, (22-1), March 1996.

[9] B. Chen, H. M. Wang, L. S. Lee, "Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 5, July 2002.